High-dimensional Bayesian network inference from systems genetics data using genetic node ordering

Lingfei Wang[1], Pieter Audenaert[2,3] and Tom Michoel[1,4*]

December 22, 2018

[1] Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK
[2] Ghent University - imec, IDLab, Technologiepark 15, 9052 Ghent, Belgium
[3] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium
[4] Computational Biology Unit, Department of Informatics, University of Bergen, PO Box 7803, 5020 Bergen, Norway
* Corresponding author, email: tom.michoel@uib.no

Abstract

Studying the impact of genetic variation on gene regulatory networks is essential to understand the biological mechanisms by which genetic variation causes variation in phenotypes. Bayesian networks provide an elegant statistical approach for multi-trait genetic mapping and modelling causal trait relationships. However, inferring Bayesian gene regulatory networks from high-dimensional genetics and genomics data is challenging, because the number of possible networks scales super-exponentially with the number of nodes, and the computational cost of Markov chain Monte Carlo (MCMC) sampling methods quickly becomes prohibitive. We propose an alternative method to infer high-quality Bayesian gene networks that easily scales to thousands of genes. Our method first reconstructs a total node ordering by conducting pairwise causal inference tests between genes, which then allows to infer a Bayesian network via a series of penalized regressions, one for each gene. We demonstrate using simulated and real systems genetics data that this results in a Bayesian network with equal, and sometimes better, likelihood than the traditional MCMC methods, while having a significantly higher overlap with groundtruth networks and being orders of magnitude faster. Moreover our method allows for a unified false discovery rate control across genes and individual edges, and thus a rigorous and easily interpretable way for tuning the sparsity level of the inferred network. Bayesian network inference using pairwise node ordering is a highly efficient approach for reconstructing gene regulatory networks when prior information for the inclusion of edges exists or can be inferred from the available data.

1 Introduction

Complex traits and diseases are driven by large numbers of genetic variants, mainly located in non-coding, regulatory DNA regions, affecting the status of gene regulatory networks [1–5]. While important progress has been made in the experimental mapping of protein-protein and protein-DNA interactions [6–8], the cell-type specific and dynamic nature of these interactions means that comprehensive, experimentally validated, cell-type or tissue-specific gene networks are not readily available for human or animal model systems. Furthermore, knowledge of physical protein-DNA interactions does not always allow to predict functional effects on target gene expression [9]. Hence, statistical and computational methods are essential to reconstruct context-specific, causal, trait-associated networks by integrating genotype and gene, protein and/or metabolite expression data from a large number of individuals segregating for the trait of interest [1–3].

Bayesian networks are a popular and powerful approach for modelling gene networks and causal relationships more generally [10–12]. They naturally extend linear models for mapping the genetic architecture of complex traits to the modelling of conditional independence and causal dependence between multiple traits, including molecular abundance traits [13–17], and have been used successfully to identify key driver genes of, for instance, type 1 diabetes [18], Alzheimer disease [19, 20], temporal lobe epilepsy [21] and cardiovascular disease [22] from systems genetics data. A Bayesian gene network consists of a directed graph without cycles, which connects regulatory genes to their targets, and which encodes conditional independence between genes. The structure and model parameters of a Bayesian network are usually inferred from the data using Markov chain Monte Carlo (MCMC) methods, whereby, starting from a randomly initialized graph, random edge additions, deletions or inversions are accepted as long as they improve the likelihood of the model [10, 11]. MCMC methods have been shown to perform well using simulated genetics and genomics data [23, 24], but their computational cost is high. Because the number of possible graphs scales super-exponentially with the number of nodes, Bayesian gene network inference with MCMC methods is feasible for systems of at most a few hundred genes, and usually requires a preliminary dimension reduction step, such as filtering or clustering genes based on their expression profiles [14, 19, 20, 22]. Modern sequencing technologies however generate transcript abundance data for ten-thousands of coding and non-coding genes, and large sample sizes mean that ever more of those are detected as variable across individuals [25–27]. Moreover, to explain why genetic associations are spread across most of the genome, a recently proposed "omnigenic" model of complex traits posits that gene regulatory networks are sufficiently interconnected such that all genes expressed in a disease or trait-relevant cell or tissue type affect the functions of core trait-related genes [5]. The limitations of current Bayesian gene network inference methods mean that this model can be neither tested nor accomodated. Hence there is a clear and unmet need to infer Bayesian networks from very high-dimensional systems genetics data.

Here we propose a novel method to infer high-quality causal gene networks that scales easily to ten-thousands of genes. Our method is based on the fact that if an ordering of nodes is given, such that the parents of any node must be a subset of the predecessors of that node in the given ordering, then Bayesian network inference reduces to a series of individual (penalized) regressions, one for each node [11, 28]. While reconstructing a node ordering is challenging in most application domains, *pairwise* comparisons between nodes can sometimes be obtained. If prior information is available for the likely inclusion of every edge, our method ranks edges according to the strength of their prior evidence (e.g. p-value) and incrementally assembles them in a directed acyclic graph, which defines a node ordering, by skipping edges that would introduce a cycle. Prior pairwise knowledge in systems biology includes the existence of TF binding motifs [29], or known protein-DNA and protein-protein

interactions [30, 31], and those have been used together with MCMC methods in Bayesian network inference previously [32, 33].

In systems genetics, where genotype and gene expression data are available for the same samples, instead of using external prior interaction data, pairwise causal inference methods can be used to estimate the likelihood of a causal interaction between every pair of genes [34–40]. To accomodate the fact that the same gene expression data is used to derive the node ordering and subsequent Bayesian network inference, we propose a novel generative model for genotype and gene expression data, given the structure of a gene regulatory graph, whose log-likelihood decomposes as a sum of the standard log-likelihood for observing the expression data and a term involving the pairwise causal inference results. Our method can then be interpreted as a straightforward greedy optimization of the posterior log-likelihood of this generative model.

2 Methods

2.1 An algorithm for the inference of Bayesian gene networks from systems genetics data

To allow the inference of Bayesian gene networks from high-dimensional systems genetics data, we developed a method that exploits recent algorithmic developments for highly efficient mapping of expression quantitative trait loci (eQTL) and pairwise causal interactions. A general overview of the method is given here, with concrete procedures for every step detailed in subsequent sections below.

A. eQTL mapping When genome-wide genotype and gene expression data are sampled from the same unrelated individuals, fast matrix-multiplication based methods allow for the efficient identification of statistically significant eQTL associations [41–44]. Our method takes as input a list of genes, and for every gene its most strongly associated eQTL (Figure 1A). Typically only *cis*-acting eQTLs (i.e. genetic variants located near the gene of interest) are considered for this step, but this is not a formal requirement. Multiple genes can have the same associated eQTL, and genes without significant eQTL can be included as well, although these will only be allowed to have incoming edges in the resultant Bayesian networks.

B. Pairwise causal ordering Given a set of genes and their respective eQTLs, pairwise causal interactions between all genes are inferred using the eQTLs as instrumental variables (Figure 1B). While there is a great amount of literature on this subject (cf. Introduction), only two stand-alone software packages are readily available: CIT [39] and Findr [40]. In our experience, only Findr is sufficiently efficient to test for causality between millions of gene pairs.

C. Genetic node ordering In Section 2.3 we introduce a generative probabilistic model for jointly observing eQTL genotypes and gene expression levels given the structure of a gene regulatory network. In this model, the posterior log-likelihood of the network given the data decomposes as a sum of two terms, one measuring the fit of the undirected network to the correlation structure of the gene expression data, and the other measuring the fit of the edge directions to the pairwise causal interactions inferred using the eQTLs as instrumental variables. The latter is optimized by a maximal directed

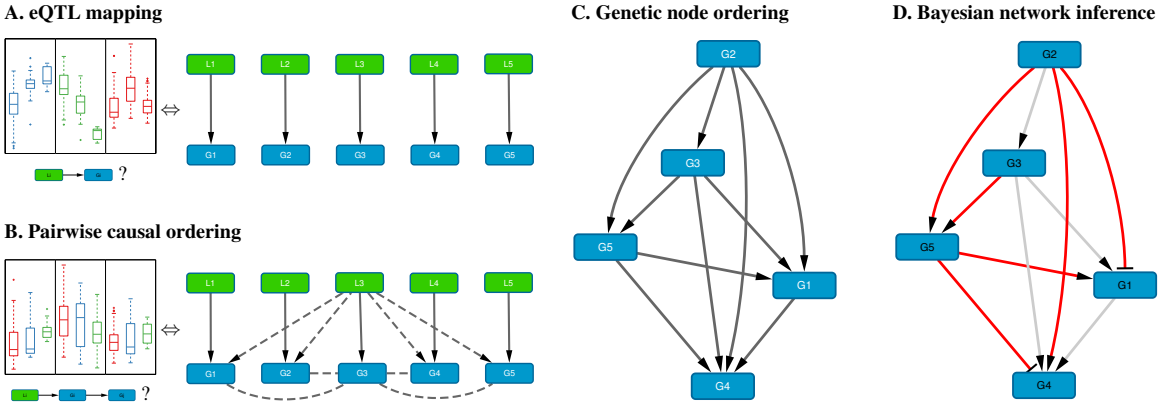

Figure 1: Schematic overview of the method. **A.** For each gene G_i, the *cis*-eQTL L_i whose genotype explains most of the variation in G_i expression is calculated; shown on the left are typical eQTL associations for three genes (colored blue, green and red) where each box shows the distribution of expression values for samples having a particular genotype for that gene's eQTL. **B.** Pairwise causal inference is carried out which considers in turn each gene G_i and its eQTL L_i to calculate the likelihood of this gene being causal for all others; shown on the left is a typical example where an eQTL L_i is associated with expression of G_i (red) and with expression of a correlated gene G_j (blue), but not with expression of G_j adjusted for G_i (green), resulting in a high likelihood score for the causal ordering $G_i \to G_j$. **C.** A total ordering is derived from the pairwise causal interactions, which can be represented as a maximal directed acyclic graph having the genes as its nodes. **D.** Variable selection is used to determine a sparse Bayesian gene network, which must be a sub-graph of the total ordering graph (red edges, Bayesian network; gray edges, causal orderings deemed not significant or indirect by the variable selection procedure); the signs of the maximum-likelihood linear regression coefficients determine whether an edge is activating (arrows) or repressing (blunt tips).

acyclic graph (DAG) or total node ordering, which we term "genetic node ordering" in reference to the use of individual-level genotype data to orient pairs of gene expression traits (Figure 1C).

D. Bayesian network inference The genetic node ordering fixes the directions of the Bayesian network edges. Variable selection methods are then used to determine the optimal sparse representation of the inverse covariance matrix of the gene expression data by a subgraph of the total ordering DAG (Figure 1D). In this paper, we use both a simple truncation of the pairwise interaction scores in the complete DAG, and multi-variate, L1-penalized lasso regression [45] to select upstream regulators for every gene. Given a sparse DAG, maximum-likelihood linear regression is used to determine the input functions and whether an edge is activating or repressing.

2.2 Bayesian network model with prior edge information

A Bayesian network with n nodes (random variables) is defined by a DAG $\mathscr{G}$ such that the joint distribution of the variables decomposes as

$$p(x_1,\ldots,x_n \mid \mathscr{G}) = \prod_{j=1}^{n} p\big(x_j \mid \{x_i \colon i \in \mathrm{Pa}_j\}\big), \tag{1}$$

where Pa_j denotes the set of parent nodes of node j in the graph $\mathscr{G}$. We only consider linear Gaussian networks [11], where the conditional distributions are given by normal distributions whose means depend linearly on the parent values (see Supplementary Information).

The likelihood of observing a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ with expression levels of n genes in m independent samples given a DAG $\mathscr{G}$ is computed as

$$p\big(\mathbf{X} \mid \mathscr{G}\big) = \prod_{k=1}^{m} \prod_{j=1}^{n} p\big(x_{jk} \mid \{x_{ik} \colon i \in \mathrm{Pa}_j\}\big). \tag{2}$$

Using Bayes' theorem we can then write the likelihood of observing $\mathscr{G}$ given the data $\mathbf{X}$, upto a normalization constant, as

$$P\big(\mathscr{G} \mid \mathbf{X}\big) \propto p\big(\mathbf{X} \mid \mathscr{G}\big) P(\mathscr{G})$$

where $P(\mathscr{G})$ is the prior probability of observing $\mathscr{G}$. Note that we use a lower-case 'p' to denote probability density functions and upper-case 'P' to denote discrete probability distributions.

Our method is applicable if pairwise prior information is available, i.e. for prior distributions satisfying

$$\log P(\mathscr{G}) \propto \sum_{(i,j) \in \mathscr{G}} f_{ij},$$

with f_{ij} a set of non-negative weights that are monotonously increasing in our prior belief that there exists a directed edge from node i to node j (e.g. $f_{ij} \propto -\log p_{ij}$, where p_{ij} is a p-value). Note that setting $f_{ij} = 0$ excludes the edge (i, j) from being present in $\mathscr{G}$.

2.3 Bayesian network model for systems genetics data

When genotype and gene expression data are available for the same samples, instrumental variable methods can be used to infer the likelihood of a causal interaction between every pair of genes [34–40]. Previously, such pairwise probabilities have been used as priors in MCMC-based Bayesian network inference [13, 23], but this is unsatisfactory, because a prior, by definition, should not be inferred from the same expression data that is used to learn the model. Other methods have addressed this by augmenting the gene network model with genotypic variables [15, 16], but this increases the size and complexity of the model even further. Here we introduce a model to use pairwise causal inference that does not suffer from these limitations.

Let $\mathscr{G}$ and $\mathbf{X}$ again be a DAG and a matrix of gene expression data for n genes, respectively, and let $\mathbf{E} \in \mathbb{R}^{n \times m}$ be a matrix of genotype data for the same samples. For simplicity we assume that each gene has one associated genotypic variable (e.g. its most significant *cis*-eQTL), but this can be extended easily to having more than one eQTL per gene or to some genes having no eQTLs. Using the rules of conditional probability, the joint probability (density) of observing $\mathbf{X}$ and $\mathbf{E}$ given $\mathscr{G}$ can be written, upto a normalization constant, as

$$p(\mathbf{X}, \mathbf{E} \mid \mathscr{G}) \propto P(\mathbf{E} \mid \mathbf{X}, \mathscr{G})\, p(\mathbf{X} \mid \mathscr{G}). \tag{3}$$

The distribution $p(\mathbf{X} \mid \mathscr{G})$ is obtained from the standard Bayesian network equations [eq. (2)], and we define the conditional probability of observing $\mathbf{E}$ given $\mathbf{X}$ and $\mathscr{G}$ as

$$P(\mathbf{E} \mid \mathbf{X}, \mathscr{G}) \propto \prod_{(i,j) \in \mathscr{G}} P(L_i \to G_i \to G_j \mid E_i, X_i, X_j), \tag{4}$$

where $E_i, X_i \in \mathbb{R}^m$ are the ith rows of $\mathbf{E}$ and $\mathbf{X}$, respectively, and $P(L_i \to G_i \to G_j \mid E_i, X_i, X_j)$ is the probability of a causal interaction from gene G_i to G_j inferred using G_i's eQTL L_i as a causal anchor. In other words, conditional on a gene-to-gene DAG $\mathscr{G}$ and a gene expression data matrix, our model assumes that it is more likely to observe genotype data that would lead to causal inferences consistent with $\mathscr{G}$ than data that would lead to inconsistent inferences. Other variations on this model can be considered as well, for instance one can include a penalty for interactions that are not present in the graph, as long as the final model can be expressed in the form

$$P(\mathbf{E} \mid \mathbf{X}, \mathscr{G}) \propto \prod_{(i,j)\in\mathscr{G}} e^{g_{ij}} \tag{5}$$

with g_{ij} monotonously increasing in the likelihood of a causal inference $L_i \to G_i \to G_j$.

Combining eqs. (3) and (5) with Bayes' theorem and a uniform prior $P(\mathscr{G}) = \text{const}$, leads to an expression of the posterior log-likelihood that is formally identical to the model with prior edge information,

$$\log P(\mathscr{G} \mid \mathbf{X}, \mathbf{E}) = \log p(\mathbf{X} \mid \mathscr{G}) + \sum_{(i,j)\in\mathscr{G}} g_{ij} + \text{const} \tag{6}$$

As before, if $g_{ij} = 0$, the edge (i, j) is excluded from being part of $\mathscr{G}$; this would happen for instance if gene i has no associated genotypic variables and consequently zero probability of being causal for any other genes given the available data. Naturally, informative pairwise graph priors of the form $\log P(\mathscr{G}) = \sum_{(i,j)\in\mathscr{G}} f_{ij}$, can still be added to the model, when such information is available.

2.4 Bayesian network parameter inference

Given a DAG $\mathscr{G}$, the maximum-likelihood parameters of the conditional distributions [eq. (1)], in the case of linear Gaussian networks, are obtained by linear regression of a gene on its parents' expression profiles (see Supplementary Information). For a specific DAG, we will use the term "Bayesian network" to refer to both the DAG itself as well as the probability distribution induced by the DAG with its maximum-likelihood parameters.

2.5 Reconstruction of the node ordering

Without further sparsity constraints in eq. (6), and again assuming for simplicity that each gene has exactly one eQTL, the log-likelihood is maximized by a maximal DAG with $n(n-1)/2$ edges. Such a DAG $\mathscr{G}$ defines a node ordering $\prec$ where $i \prec j \Leftrightarrow (i, j) \in \mathscr{G}$. Standard results in Bayesian network theory show that for a linear Gaussian network, the likelihood function (2) is invariant under arbitrary changes of the node ordering (see [11] and Supplementary Information). Hence to maximize eq. (6) we need to find the node ordering or DAG which maximizes the term $\sum_{(i,j)\in\mathscr{G}} g_{ij}$. Finding the maximum-weight acyclic subgraph is an NP-hard problem with no known polynomial approximation algorithms with a strong guaranteed error bound [46, 47]. We therefore employed a greedy algorithm, where given n genes and the log-likelihood g_{ij} of regulation between every pair of them, we first rank the regulations according to their likelihood. The regulations are then added to an empty network one at a time starting from the most probable one, but avoiding those that would create a cycle, until a maximal DAG with $n(n-1)/2$ edges is obtained. Other edges are assigned probability 0 to indicate exclusion. Maximal DAG reconstruction was implemented in `Findr` [48] as the command `netr_one_greedy`, with the *vertex-guided* algorithm for cycle detection [49].

2.6 Causal inference of pairwise gene regulations

We used Findr 1.0.6 (pij_gassist function) [48] to perform causal inference of gene regulatory interactions based on gene expression and genotype variation data. For every gene, its strongest *cis*-eQTL was used as a causal anchor to infer the probability of regulation between that gene and every other gene. Findr outputs posterior probabilities P_{ij} (i.e. one minus local FDR), which served directly as weights in model (6), i.e. we set $g_{ij} = \log P_{ij}$. To verify the contribution from the inferred pairwise regulations, we also generated random pairwise probability matrices which were treated in the same way as the informative ones in the downstream analyses.

2.7 Findr and random Bayesian networks from complete node orderings

The node ordering reconstruction removes less probable, cyclic edges, and results in a maximal, weighted DAG $\mathscr{G}$ with edge weights $P_{ij} = e^{g_{ij}}$. We term these weighted, complete DAGs as *findr* or *random Bayesian networks*, depending on the pairwise information used. A significance threshold can be further applied on the continuous networks, so as to convert them to binary Bayesian networks at any desired sparsity level.

2.8 Lasso-findr and lasso-random Bayesian networks using penalized regression on ordered nodes

To infer a more refined sparse Bayesian network from a maximal DAG, we performed hypothesis testing for every gene on whether each of its predecessors (in findr or random Bayesian network) is a regulator, using L1-penalized lasso regression [45] with the lassopv package [50] (see Supplementary Information). We calculated for every regulator the p-value of the critical regularization strength when the regulator first becomes active in the lasso path. This again forms a continuous Bayesian network in which smaller p-values indicate stronger significance. These Bayesian networks were termed the *lasso-findr* and *lasso-random Bayesian networks*.

2.9 MCMC-based bnlearn-hc and bnlearn-fi Bayesian networks from package bnlearn

For comparison with the traditional MCMC-based Bayesian network inference, we applied the hc function of the R package bnlearn [51], using the Akaike information criterion (AIC) penalty to enforce sparsity. This algorithm starts from a random Bayesian network and iteratively performs greedy revisions on the Bayesian network to reach a local optimum of the penalized likelihood function. Since the log-likelihood is equivalent to minus the average (over nodes) log unexplained variance (see Supplementary Information), which diverges when the number of regulators exceeds the number of samples, we enforced the number of regulators for every gene to be smaller than 80% of the number of samples. For each AIC penalty, one hundred random restarts were carried out and only the network with highest likelihood score was selected for downstream analyses. These Bayesian networks were termed the *bnlearn-hc* Bayesian networks.

For comparison with the constraint based Bayesian network inference (e.g. [52]), we applied the fast.iamb function of the R package bnlearn [51], using nominal type I error rate. These Bayesian networks were termed the *bnlearn-fi* Bayesian networks.

To account for the role and information of cis-eQTLs on gene expression, we also included the strongest cis-eQTL of every gene in the bnlearn based network reconstructions, for an approach similar to [15, 16]. Cis-eQTLs are only allowed to have out-going edges, using the blacklist function in bnlearn. We then removed cis-eQTL nodes from the reconstructed networks, resulting in Bayesian gene networks termed *bnlearn-hc-g* and *bnlearn-fi-g* respectively.

2.10 Evaluation of false discovery control in network inference

An inconsistent false discovery control (FDC) reduces the overall accuracy of the reconstructed network [50]. We empirically evaluated the FDC using a linearity test on genes that are both targets and regulators. The linearity test assesses whether the number of false positive regulators for each gene increases linearly with the number of candidate regulators, a consequence of consistent FDC. The top 5% predictions were discarded to remove genuine interactions. See [50] for method details.

2.11 Precision-recall curves and points

We compared reconstructed Bayesian networks with gold standards using precision-recall curves and points, for continuous and binary networks respectively. For Geuvadis datasets, we only included regulator and target genes that are present in both the transcriptomic dataset and the gold standard.

2.12 Assessment of predictive power for Bayesian networks

To assess the predictive power of different Bayesian network inference methods, we used five-fold cross-validation to compute the training and testing errors from each method, in terms of the root mean squared error (rmse) and mean log squared error (mlse) across all genes in all testing data (Algorithm 1). For continuous Bayesian networks from non-bnlearn methods, we applied different significance thresholds to obtain multiple binary Bayesian networks that form a curve of prediction errors.

Algorithm 1 Cross-validation of predictive power for Bayesian networks

Require: $M \in R^{n \times m}$ as matrix of normalized expression,
$B(m) \in R^{n \times n}$ as function to infer binary Bayesian network from expression matrix m,
$s(\hat{y}, y)$ as score function (rmse or mlse) of predicted expression $\hat{y}$ given true expression y.

function CROSS-VALIDATION(M, B, s)
  $train\_score, test\_score \leftarrow 0$
  for $i \leftarrow 1$ to 5 **do**
    $train, test \leftarrow$ Random cross-validation split i of training & test data from M
    $\mathscr{G} \leftarrow B(train)$
    for $j \leftarrow 1$ to n **do**
      $model \leftarrow$ Fitted linear model to predict $train_j$ with $train_{\mathscr{G}_{\cdot,j}}$
      $train\_score \leftarrow train\_score + s(model(train_{\mathscr{G}_{\cdot,j}}), train_j)$
      $test\_score \leftarrow test\_score + s(model(test_{\mathscr{G}_{\cdot,j}}), test_j)$
  $train\_score \leftarrow train\_score/5n$
  $test\_score \leftarrow test\_score/5n$
  return $train\_score, test\_score$

2.13 Data

We used the following datasets to infer and evaluate Bayesian gene networks:

- The DREAM 5 Systems Genetics challenge A (DREAM) provided a unique testbed for network inference methods that utilize genetic variations in a population (`https://www.synapse.org/\#!Synapse:syn2820440/wiki/`). The DREAM challenge included 15 simulated datasets of expression levels of 1000 genes and their best eQTL variations. To match the high-dimensional property of real datasets where the number of genes exceeds the number of individuals, we analyzed datasets 1, 3, and 5 with 100 individuals each. Around 25% of the genes within each dataset had a cis-eQTL, defined in DREAM as directly affecting the expression level of the corresponding gene. Since the identity of cis-eQTLs is not revealed, we used kruX [53] to identify them, allowing for one false discovery per dataset. The DREAM challenge further provides the groundtruth network for each dataset, varying from around 1000 to 5000 interactions.
- The Geuvadis consortium is a population study providing RNA sequencing and genotype data of lymphoblastoid cell lines in 465 individuals. We obtained gene expression levels and genotype information, as well as the eQTL mapping from the original study [54]. We limited our analysis to 360 European individuals, and after quality control, a total of 3172 genes with significant cis-eQTLs remained. To validate the inferred gene regulatory networks from the Geuvadis dataset, we obtained three groundtruth networks: (1) differential expression data from siRNA silencing experiments of transcription-associated factors (TFs) in a lymphoblastoid cell line (GM12878) [55]; (2) DNA-binding information of TFs in the same cell line [55]; (3) the filtered proximal TF-target network from [7]. The Geuvadis dataset overlapped with 6,790 target genes, and 6 siRNA-targeted TFs and 20 DNA-binding TFs in groundtruth 1 and 2, respectively, and with 7,000 target genes and 14 TFs in groundtruth 3.

We preprocessed all expression data by converting them to a standard normal distribution separately for each gene, as explained in [48].

3 Results

3.1 Lasso-findr Bayesian networks correctly control false discoveries

We inferred findr and lasso-findr Bayesian networks for the DREAM datasets, using `Findr` and `lassopv` respectively (Methods). The `Findr` method predicts targets for each regulator using a local FDR score [56] which allows false discovery control (FDC) for either the entire regulator-by-target matrix, or for a specific regulator of interest [35, 48]. However, the enforcement of a gene ordering/Bayesian network partly broke the FDC, as seen from the linearity test (Methods) in Figure 2**A**. By performing an extra lasso regression on top of the acyclic findr network, proper FDC was restored in the lasso-findr Bayesian network (Figure 2**B**, Supplementary Figure S1).

In contrast, MCMC-based bnlearn-hc Bayesian networks (Methods), inferred from multiple DREAM datasets and for a spectrum of network sparsities (AIC penalty strengths from 8 to 12 in steps of 0.5), displayed a highly skewed in-degree distribution, with most genes having few regulators, but several with near 80 regulators each, i.e. the maximum allowed (Figure 2**C**, Supplementary Figure S2). This indicates that MCMC-based Bayesian networks lack a unified FDR control, i.e. that each gene

retained incoming interactions at different FDR levels. We believe this is due to the log-likelihood score function employed by bnlearn-hc. Since the log-likelihood corresponds to the average logarithm of the unexplained variance, this score intrinsically tends to focus on the explanation of variances from a few variables/genes, especially in high-dimensional settings where this can lead to arbitrarily large score values (see Supplementary Information). Using the total proportion of explained variance as the score may spread regulations over more target genes, but this score is not implemented in bnlearn.

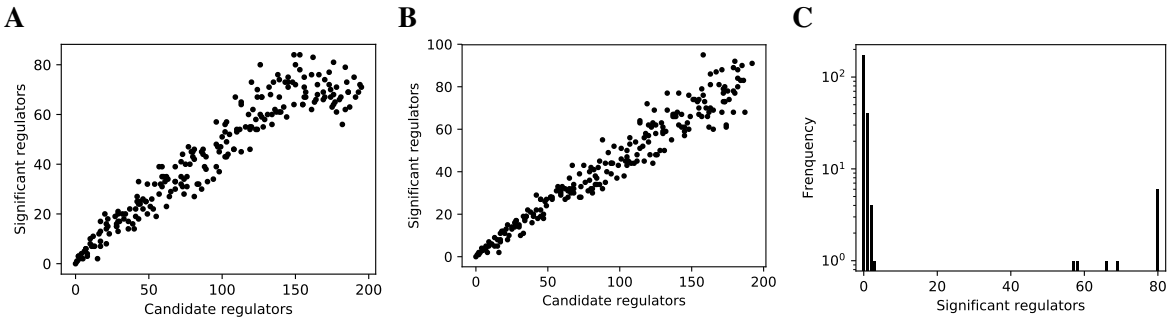

Figure 2: False discovery controls of different Bayesian networks. (**A, B**) The linearity test of findr (**A**) and lasso-findr (**B**) Bayesian networks at 10,000 significant interactions on DREAM dataset 1. (**C**) The histogram of significant regulator counts for each target gene in the bnlearn-hc Bayesian network with AIC penalty 8 on DREAM dataset 1.

Constraint-based bnlearn-fi Bayesian networks (Methods) did not allow for unbiased FDC either, as they do not have a fully adjustable sparsity level. We varied its "nominal type I error rate" from 0.001 to 0.2, but the number of significant interactions varied very little on DREAM dataset 1 (Supplementary Figure S3).

Incorporating genotypic information in MCMC-based (bnlearn-hc-g) or constraint-based (bnlearn-fi-g) Bayesian networks did not resolve these issues, as the problems of lacking FDC and oversparsity persisted (Supplementary Figure S4, Supplementary Figure S5).

3.2 Findr and lasso Bayesian networks recover genuine interactions more accurately than MCMC or constraint-based networks

We compared the inferred Bayesian networks from all methods against the groundtruth network of the DREAM challenge. We drew precision-recall (PR) curves, or points for the binary Bayesian networks from bnlearn-based methods. As shown in Figure 3, the findr, lasso-findr, and lasso-random Bayesian networks were more accurate predictors of the underlying network structure. The inclusion of genotypic information improved the precision of bnlearn methods, but it remained less optimal than findr and lasso-based Bayesian networks.

3.3 Findr and lasso Bayesian networks obtain superior predictive performances

We validated the predictive performances of all networks in the structural equation context (see Supplementary Information). Under 5-fold cross validation, a linear regression model for each gene on its parents is trained based on the Bayesian network structure inferred from each training set, to predict expression levels of all genes in the test set (Methods). Predictive errors were measured in terms of

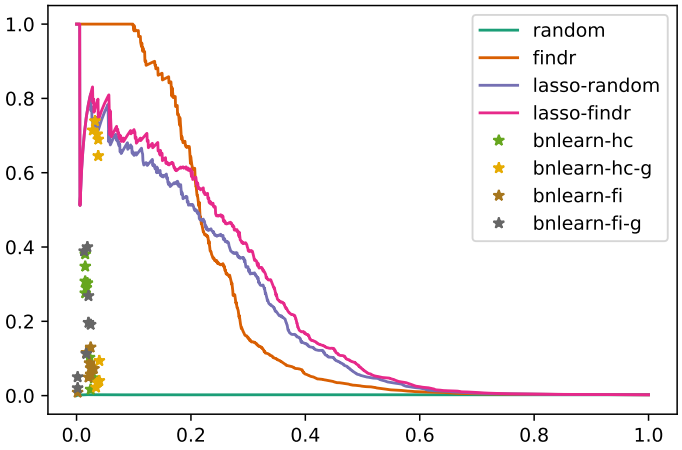

Figure 3: Precision-recall curves/points of reconstructed Bayesian networks for DREAM dataset 1.

root mean squared error (rmse) and mean log squared error (mlse; the score optimized by bnlearn-hc). The findr Bayesian network explained the highest proportion of expression variation ($\approx 2\%$) in the test data and identified the highest number of regulations (200 to 300), with runners up from lasso-based networks ($\approx 1\%$ variation, 50 regulations, Figure 4). The explained variance by findr and lasso networks grew to $\approx 10\%$ when more samples were added (DREAM dataset 11 with 999 samples, Supplementary Figure S6). Training errors did not show overfitting of predictive performances in the test data (Supplementary Figure S7).

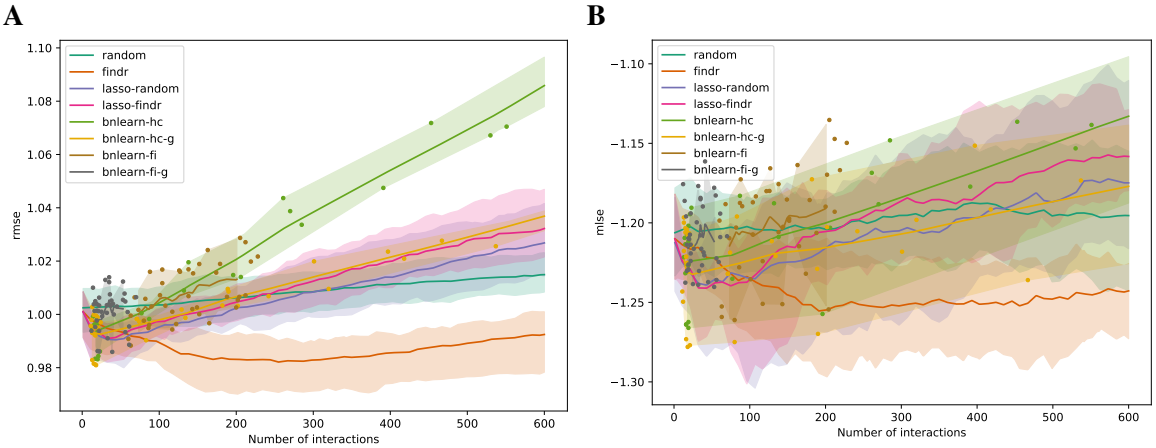

Figure 4: The root mean squared error (rmse, **A**) and mean log squared error (mlse, **B**) in test data are shown as functions of the numbers of predicted interactions in five-fold cross validations using linear regression models. Shades and lines indicate minimum/maximum values and means respectively. Root mean squared errors greater than 1 indicate over-fitting. DREAM dataset 1 with 100 samples was used.

3.4 Lasso Bayesian networks do not need accurate prior gene ordering

Interestingly, the performance of lasso-based networks did not depend strongly on the prior ordering, as shown in the comparisons between lasso-findr and lasso-random in Figure 3, Figure 4, and Supplementary Figure S7. Further inspections revealed a high overlap of top predictions by lasso-findr and lasso-random Bayesian networks, particularly among their true positives (Figure 5). This allows us to still recover genuine interactions even if the prior gene ordering is not fully accurate.

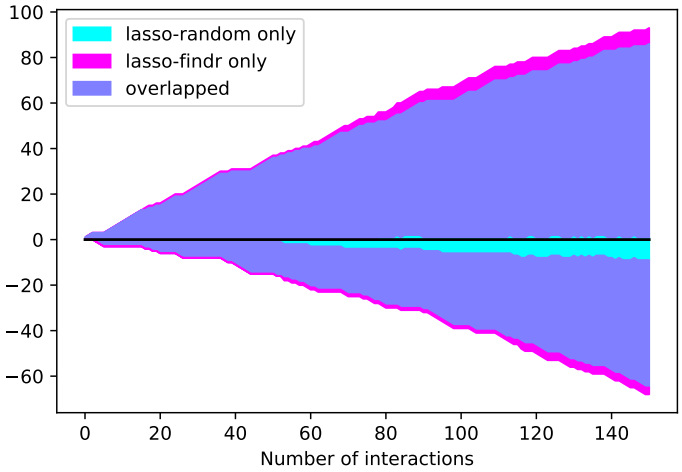

Figure 5: The numbers of overlap and unique interactions (y axis) predicted by lasso-findr and lasso-random Bayesian networks as functions of the number of significant interactions in each network (x axis), on DREAM dataset 1. Positive and negative directions in y correspond to true and false positive interactions according to the gold standard.

3.5 Lasso Bayesian networks mistaken confounding as false positive interactions

We then tried to understand the differences between lasso and Findr based Bayesian networks, by comparing three types of gene relations in DREAM dataset 1, both among genes with a cis-eQTL in Figure 6**A**, and when also including genes without any cis-eQTL as only targets in Figure 6**B**. Both findr and lasso-findr showed good sensitivity for the genuine, direct interactions. However, when two otherwise independent genes are directly confounded by another gene, lasso tends to produce a false positive interaction, but not findr. As expected, to achieve optimal predictive performance, lasso regression cannot distinguish the confounding by a gene that is either unknown or ranked lower in the DAG.

3.6 Findr and lasso Bayesian network inference is highly efficient

The findr and lasso Bayesian networks required much less computation time compared to the bnlearn Bayesian networks, therefore allowing them to be applied on much larger datasets. To infer a Bayesian network of 230 genes from 100 samples in DREAM dataset 1, Findr required less than a second, lassopv around a minute, but bnlearn Bayesian networks took half an hour to half a day

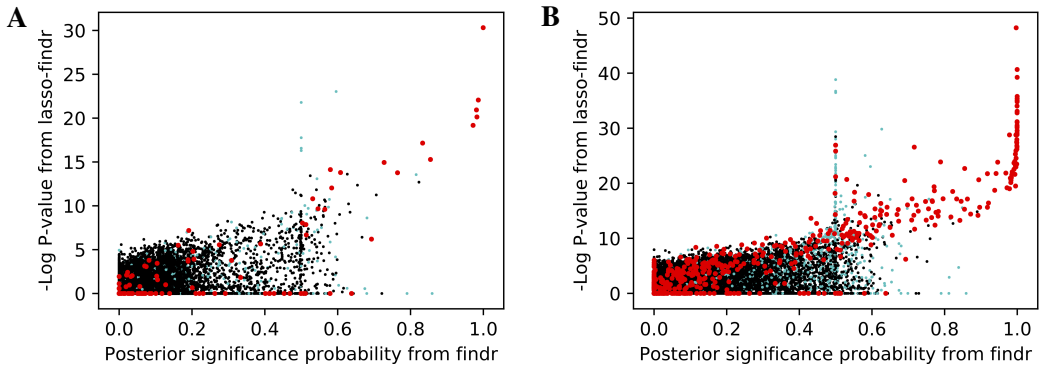

Figure 6: The significance score of findr (posterior probability; x-axis) and in lasso-findr (-log P-value; y-axis) for direct true interactions (red), directly confounded gene pairs (cyan), and other, unrelated gene pairs (black) on DREAM dataset 1; in **A**, only genes with cis-eQTLs are considered as regulator or target, whereas in **B** targets also include genes without cis-eQTLs. Higher scores indicate stronger significances for the gene pair tested.

(Table 1). Moreover, since bnlearn only produces binary Bayesian networks, multiple recomputation is necessary to acquire the desired network sparsity.

Table 1: Timings for different Bayesian network inference methods/programs. Times for bnlearn methods depend on parameter settings (e.g. nominal FDR and AIC penalty), and take longer (approx. 8 times) with genotypes included. Times for bnlearn-hc include 10 random restarts.

Dataset	Samples	Genes	findr	lassopv	bnlearn-hc	bnlearn-fi
DREAM	100	230	$<$1sec	$\approx$1min	$\geq$10hr	$\geq$30min
Geuvadis	360	3172	$<$1min	$\approx$10hr	-	-

3.7 Results on the Geuvadis dataset reaffirm conclusions from simulated data

To test whether the results from the DREAM data also hold for real data, we inferred findr and lasso-findr Bayesian networks from the Geuvadis data using both real and random causal priors (see Methods); MCMC-based bnlearn network inference was attempted, but none of the restarts could complete within 1000 minutes.

Lasso-findr Bayesian networks were previously shown to provide ideal FDR control on this dataset [50], whereas findr Bayesian networks did not obtain a satisfying FDR control (Supplementary Figure S8). We believe this is due to the reconstruction of a prior node ordering, which interferes with the FDR control in pairwise causal inference. On the other hand, and again consistent with the DREAM data, findr Bayesian networks obtained superior results for the recovery of known transcriptional regulatory interactions inferred from ChIP-sequencing data (Figure 7A,B); neither method predicted TF targets inferred from siRNA silencing with high scores or accuracy better than random (Figure 7C).

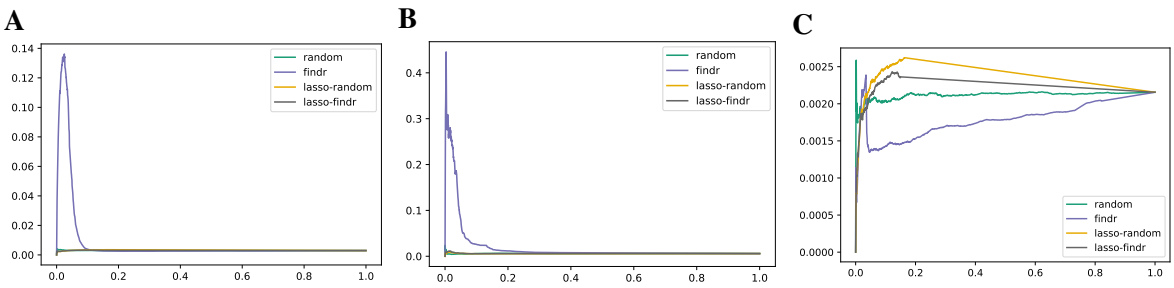

Figure 7: Precision-recall curves for Bayesian networks reconstructed from the Geuvadis dataset for three groundtruth networks: DNA-binding of 20 TFs in GM12878 (**A**), DNA-binding of 14 TFs in 5 ENCODE cell lines (**B**), and siRNA silencing of 6 TFs in GM12878 (**C**).

Comparisons on the predictive power yielded results similar with the DREAM datasets, where predictive scores were again hardly able to distinguish network directions.

4 Discussion

The inference of Bayesian gene regulatory networks for mapping the causal relationships between thousands of genes expressed in any given cell type or tissue is a challenging problem, due to the computational complexity of standard MCMC sampling methods. Here we have introduced an alternative method, which first reconstructs a total topological ordering of genes, and then infers a sparse maximum-likelihood Bayesian network using penalized regression of every gene on its predecessors in the ordering. Our method is applicable when pairwise prior information is available or can be inferred from auxiliary data, such as genotype data. Our evaluation of the method using simulated genotype and gene expression data from the DREAM5 competition, and real data from human lymphoblastoid cell lines from the GEUVADIS consortium, revealed several lessons that we believe to be generalizable.

A major disadvantage of MCMC methods, irrespective of their computational cost, was their overfitting of the expression profiles of a very small number of target genes. In high-dimensional settings where the number of genes far exceeds the number of samples, the expression profile of any one of them can be regressed perfectly (i.e. with zero residual error) on any linearly independent subset of variables, and this causes the log-likelihood to diverge. Even when the number of parents per gene was restricted to less than the number of samples, it remained the case that at any level of network sparsity, the divergence of the log-likelihood with decreasing residual variance of even a single gene resulted in MCMC networks where most genes had either the maximum number of parents, or no parents at all. Restricting the maximum number of parents to an artificially small level can circumvent this problem, but will also distort the network topology, particularly by truncating the in-degree distribution, and therefore predict a biased gene regulatory network. Optimizing the total amount of variance explained, rather than log-likelihood, might overcome this problem. This, however, is not available yet in bnlearn.

Our method assembles a global Bayesian network from pairwise causal relationships inferred using instrumental variable methods. We considered two variants of the method: one where the pairwise relations were truncated directly to form a sparse DAG, and one where an additional L1-penalized lasso regression step was used to enforce sparsity. The lasso step was introduced for two reasons.

First, pairwise relations do not distinguish between direct or indirect interactions and do not account for the possibility that a true relation may only explain a small proportion of target gene variation (e.g. when the target has multiple inputs). We hypothesized that adding a multi-variate lasso regression step could address these limitations. Second, truncating pairwise relations directly results in non-uniform false discovery rates for the retained interactions, due to each gene starting with a different number of candidate parents in the pairwise node ordering. As we showed in this paper and our previous work [50], a model selection p-value derived from lasso regression can control the FDR uniformly for each potential regulator of each target gene, resulting in an unbiased sparse DAG.

Despite these considerations, the 'naive' procedure of truncating the original pairwise causal probabilities resulted in Bayesian networks with better overlap with groundtruth networks of known transcriptional interactions, in both simulated and real data. We believe this is due to the lack of any instrumental variables in lasso regression, which makes it hard to dissociate true causal interactions from hidden confounding. Indeed, it is known that if there are multiple strongly correlated predictors, lasso regression will randomly select one of them [57], whereas in the present context it would be better to select the one that has the highest prior causal evidence. In a real biological system, findr networks and the use of instrumental variables may therefore be more robust than lasso regression, particularly in the presence of hidden confounders. We also note that the deviation from uniform FDR control for the naive truncation method was not huge and only affected genes with a very large number of candidate parents (Figure 2). Hence, at least in the datasets studied, adding a lasso step for better false discovery control did not overcome the limitations introduced by confounding interactions.

On the other hand, the lasso-random network uses solely transcriptomic profiles, yet provided better performance than all MCMC-based networks, including those that used genotypic information. Together with its better false discovery control, this makes the lasso-random network the ideal method for Bayesian network inference with no or limited prior information.

In addition to comparing the inferred network structure against known ground-truths, we also compared the predictive performance of the various Bayesian networks. Although findr Bayesian networks again performed best, differences with lasso-based methods were modest. As is well known, using observational data alone, Bayesian networks are only defined upto Markov equivalence [11, 12], i.e. there is usually a large class of Bayesian networks with very different topology which all explain the data equally well. Hence it comes as no surprise that the prediction accuracy in edge directions has little impact on that in expression levels. This suggests that for the task of reconstructing gene networks, Bayesian network inference should be evaluated, and maybe also optimized, at the structural rather than inferential level. This also reinforces the importance of causal inference which, although challenging both statistically and computationally, demonstrated significant improvement of the global network structure even when it was restricted to pairwise causal tests.

Most of our results were derived for simulated data from the DREAM Challenges, but were qualitatively confirmed using data from human lymphoblastoid cell lines. However, it has to be acknowledged that the human ground-truth networks are small. Because the available networks for lymphoblastoid cell lines are exclusively for TF-DNA or TF siRNA interactions, and TFs tend not to show great variation in transcriptional expression data (i.e. don't have very strongly associated eQTLs), only 6–20 TFs were common between the predicted and ground-truth networks. As such, one has to be cautious not to over-interpret results, for instance on the relative performance of findr vs. lasso-findr Bayesian networks. Much more comprehensive and accurate ground-truth networks of direct causal interactions, preferably derived from a hierachy of interventions on a much wider variety of genes and functional classes, would be required for a conclusive analysis. Emerging large-scale

perturbation compendia such as the expanded Connectivity Map, which has profiled knock-downs or over-expressions of more than 5,000 genes in a variable number of cell lines using a reduced representation transcriptome [58], hold great promise. However, the available cell lines are predominantly cancer lines, and the relevance of the profiled interactions for systems genetics studies of human complex traits and diseases, which are usually performed on primary human cell or tissue types, remains unknown.

Lastly, we note that our study has focused on ground-truth comparisons and predictive performances, but did not evaluate how well the second part of the log-likelihood, derived from the genotype data [cf. eq. (4)], was optimized. This score is never considered in MCMC-based algorithms, and hence a comparison would not be fair, and moreover, optimising it is known to be an NP-hard problem. We used a common greedy heuristic optimization algorithm, but for this particular problem, this heuristic has no strong guaranteed error bound. We intend to revisit this problem, and investigate whether other graph-theoretical algorithms, perhaps tailored to specific characteristics of pairwise interactions inferred from systems genetics data, are able to improve on the greedy heuristic.

To conclude, Bayesian network inference using pairwise node ordering is a highly efficient approach for reconstructing gene regulatory networks from high-dimensional systems genetics data, which outperforms traditional MCMC-based methods by assembling pairwise causal inference results in a global causal network, and which is sufficiently flexible to integrate other types of pairwise prior data when they are available.

Funding

This work has been supported by the BBSRC (grant numbers BB/J004235/1 and BB/M020053/1).

References

[1] Matthew V Rockman. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature*, 456(7223):738–744, 2008.

[2] E E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461:218–223, 2009.

[3] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, 2014.

[4] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16:197–212, 2015.

[5] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[6] Albertha JM Walhout. Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping. *Genome Research*, 16(12):1445–1454, 2006.

[7] M.B. Gerstein, A. Kundaje, M. Hariharan, S.G. Landt, K.K. Yan, C. Cheng, X.J. Mu, E. Khurana, J. Rozowsky, R. Alexander, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

[8] Katja Luck, Gloria M Sheynkman, Ivy Zhang, and Marc Vidal. Proteome-scale human interactomics. *Trends in biochemical sciences*, 42(5):342–354, 2017.

[9] Darren A Cusanovich, Bryan Pavlovic, Jonathan K Pritchard, and Yoav Gilad. The functional consequences of variation in transcription factor binding. *PLoS Genetics*, 10(3):e1004226, 2014.

[10] N Friedman, M Linial, I Nachman, and D Pe'er. Using Bayesian networks to analyze expression data. *J Comput Biol*, 7:601–620, 2000.

[11] D Koller and N Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[12] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[13] J Zhu, P Y Lum, J Lamb, D GuhaThakurta, S W Edwards, R Thieringer, J P Berger, M S Wu, J Thompson, A B Sachs, and E E Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, 105:363–374, 2004.

[14] J Zhu, B Zhang, E N Smith, B Drees, R B Brem, L Kruglyak, R E Bumgardner, and E E Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40:854–861, 2008.

[15] Elias Chaibub Neto, Mark P Keller, Alan D Attie, and Brian S Yandell. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, 4(1):320, 2010.

[16] Rachael S Hageman, Magalie S Leduc, Ron Korstanje, Beverly Paigen, and Gary A Churchill. A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics*, 187(4):1163–1170, 2011.

[17] Marco Scutari, Phil Howell, David J Balding, and Ian Mackay. Multiple quantitative trait analysis using Bayesian networks. *Genetics*, 198(1):129–137, 2014.

[18] E. E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, and R. Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, 6:e107, 2008.

[19] B. Zhang, C. Gaiteri, L. G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu, and V. Emilsson. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, 153(3):707–720, Apr 2013.

[20] Noam D Beckmann, Wei-Jye Lin, Minghui Wang, Ariella T Cohain, Pei Wang, Weiping Ma, Ying-Chih Wang, Cheng Jiang, Mickael Audrain, Phillip Comella, et al. Multiscale causal network models of Alzheimer's disease identify VGF as a key regulator of disease. *bioRxiv*, page 458430, 2018.

[21] Michael R Johnson, Jacques Behmoaras, Leonardo Bottolo, Michelle L Krishnan, Katharina Pernhorst, Paola L Meza Santoscoy, Tiziana Rossetti, Doug Speed, Prashant K Srivastava, Marc Chadeau-Hyam, et al. Systems genetics identifies sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. *Nature communications*, 6:6031, 2015.

[22] H. Talukdar, H Foroughi Asl, R. Jain, R. Ermel, A. Ruusalepp, O. Franzén, B. Kidd, B. Readhead, C. Giannarelli, T. Ivert, J. Dudley, M. Civelek, A. Lusis, E. Schadt, J. Skogsberg, T. Michoel, and J.L.M Björkegren. Cross-tissue regulatory gene networks in coronary artery disease. *Cell Systems*, 2:196–208, 2016.

[23] Jun Zhu, Matthew C Wiener, Chunsheng Zhang, Arthur Fridman, Eric Minch, Pek Y Lum, Jeffrey R Sachs, and Eric E Schadt. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol*, 3(4):e69, 2007.

[24] Shinya Tasaki, Ben Sauerwine, Bruce Hoff, Hiroyoshi Toyoshiba, Chris Gaiteri, and Elias Chaibub Neto. Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. *Genetics*, 199(4):973–989, 2015.

[25] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501:506–511, 2013.

[26] O Franzén, R Ermel, A Cohain, N Akers, A Di Narzo, H Talukdar, H Foroughi Asl, C Giambartolomei, J Fullard, K Sukhavasi, S Köks, L-M Gan, C Gianarelli, J Kovacic, C Betsholtz, B Losic, T Michoel, K Hao, P Roussos, J Skogsberg, A Ruusalepp, E Schadt, and J Björkegren. Cardiometabolic risk loci share downstream *cis* and *trans* genes across tissues and diseases. *Science*, 2016.

[27] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[28] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[29] H J Bussemaker, B C Foat, and L D Ward. Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu Rev Biophys Biomol Struct*, 36:329–347, 2007.

[30] J Ernst, Q K Beg, K A Kay, G Balázsi, Z N Oltvai, and Z Bar-Joseph. A semi-supervised method for predicting transcription factor - gene interactions in *Escherichia coli*. *PloS Comp Biol*, 4:e1000044, 2008.

[31] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, 2013.

[32] Adriano V Werhli and Dirk Husmeier. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6(1):15, 2007.

[33] Sach Mukherjee and Terence P Speed. Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318, 2008.

[34] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

[35] Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8(10):R219, 2007.

[36] Joshua Millstein, Bin Zhang, Jun Zhu, and Eric E Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10(1):23, 2009.

[37] Yang Li, Bruno M Tesson, Gary A Churchill, and Ritsert C Jansen. Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in Genetics*, 26(12):493–498, 2010.

[38] Elias Chaibub Neto, Aimee T Broman, Mark P Keller, Alan D Attie, Bin Zhang, Jun Zhu, and Brian S Yandell. Modeling causality for pairs of phenotypes in system genetics. *Genetics*, 193(3):1003–1013, 2013.

[39] Joshua Millstein, Gary K Chen, and Carrie V Breton. cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*, 32:2364–2365, 2016.

[40] Lingfei Wang and Tom Michoel. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLoS Computational Biology*, 13(8):e1005703, 2017.

[41] Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[42] J. Qi, H. Foroughi Asl, J. L. M. Björkegren, and T. Michoel. kruX: Matrix-based non-parametric eQTL discovery. *BMC Bioinformatics*, 15:11, 2014.

[43] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2015.

[44] Olivier Delaneau, Halit Ongen, Andrew A Brown, Alexandre Fort, Nikolaos I Panousis, and Emmanouil T Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, 8:15452, 2017.

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

[46] Bernhard Korte and Dirk Hausmann. An analysis of the greedy heuristic for independence systems. In *Annals of Discrete Mathematics*, volume 2, pages 65–74. Elsevier, 1978.

[47] Refael Hassin and Shlomi Rubinstein. Approximations for the maximum acyclic subgraph problem. *Information processing letters*, 51(3):133–140, 1994.

[48] Lingfei Wang and Tom Michoel. Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PLOS Computational Biology*, 13(8):e1005703, August 2017.

[49] Bernhard Haeupler, Telikepalli Kavitha, Rogers Mathew, Siddhartha Sen, and Robert E. Tarjan. Incremental cycle detection, topological ordering, and strong component maintenance. *ACM Trans. Algorithms*, 8(1):3:1–3:33, January 2012.

[50] Lingfei Wang and Tom Michoel. Controlling false discoveries in Bayesian gene networks with lasso regression p-values. *arXiv:1701.07011 [q-bio, stat]*, January 2017. arXiv: 1701.07011.

[51] Marco Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(1):1–22, 2010.

[52] Markus Kalisch and Peter Bühlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[53] Jianlong Qi, Hassan Foroughi Asl, Johan Bjorkegren, and Tom Michoel. krux: matrix-based non-parametric eqtl discovery. *BMC Bioinformatics*, 15(1):11, 2014.

[54] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedlander, Peter A. C. /'t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlof, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Hasler, Ann-Christine Syvanen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigo, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 09 2013.

[55] Darren A Cusanovich, Bryan Pavlovic, Jonathan K Pritchard, and Yoav Gilad. The functional consequences of variation in transcription factor binding. *PLoS Genetics*, 10(3):e1004226, 2014.

[56] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, August 2003.

[57] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[58] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Supplementary Information

S1 Theoretical background and results

S1.1 Bayesian network primer

We collect here the minimal background on Bayesian networks necessary to make this paper self-contained. For more details and proofs of the statements below, we refer to existing textbooks, for instance [11].

A Bayesian network for a set of continuous random variables $X_1,\dots,X_n$, represented by nodes $1,\dots,n$, is defined by a DAG $\mathscr{G}$ and a joint probability density function that decomposes as in eq. (1). We are interested in linear Gaussian networks, which can be defined alternatively by the set of structural equations

$$X_j = \sum_{i\in \mathrm{Pa}_j} \beta_{ij} X_i + \varepsilon_j, \tag{S1}$$

where Pa_j is the set of parent nodes for node j in $\mathscr{G}$ and $\varepsilon_j \sim \mathscr{N}(0,\omega_j^2)$ are mutually independent normally distributed variables. The matrix $\mathbf{B} = (\beta_{ij})$, with $\beta_{kj} = 0$ for $k \notin \mathrm{Pa}_j$ can be regarded as a weighted adjacency matrix for $\mathscr{G}$. With this notation, the conditional distributions in eq. (1) satisfy

$$p\big(x_j \mid \{x_i\colon i \in \mathrm{Pa}_j\}\big) = \mathscr{N}\Big(\sum_{i\in\mathrm{Pa}_j} \beta_{ij} x_i, \omega_j^2\Big). \tag{S2}$$

The values of $\mathbf{B}$ and $\omega_1^2,\dots,\omega_n^2$ are the parameters of the Bayesian network which are to be determined along with the structure of $\mathscr{G}$. The conditional distributions (S2) result in the joint probability density function being multi-variate normal,

$$p(x_1,\dots,x_n) = \prod_{j=1}^{n} p\big(x_j \mid \{x_i\colon i\in\mathrm{Pa}_j\}\big) = \mathscr{N}(0,\Sigma)$$

with inverse covariance matrix

$$\Sigma^{-1} = (\mathbb{1} - \mathbf{B})\Omega^{-1}(\mathbb{1}-\mathbf{B})^T$$

where $\Omega = \mathrm{diag}(\omega_1^2,\dots,\omega_n^2)$. It follows that the gene expression-based term in the log-likelihood [eq. (6)] can be written as (up to an additive constant)

$$\mathscr{L}_X \equiv \log p(\mathbf{X} \mid \mathscr{G}) = \frac{m}{2}\log\det\Sigma^{-1} - \frac{1}{2}\mathrm{tr}\big(\Sigma^{-1}\mathbf{X}\mathbf{X}^T\big) \tag{S3}$$

where as before $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the data matrix for n genes in m independent samples. From these basic results, the following can be derived easily:

- For a given $\mathscr{G}$, $\mathscr{L}_X$ can also be written as

$$\mathscr{L}_X = \sum_{j=1}^{n}\left[-\frac{m}{2}\log(\omega_j^2) - \frac{1}{2\omega_j^2}\Big\|X_j - \sum_{i\in\mathrm{Pa}_j}\beta_{ij}X_i\Big\|^2\right] \tag{S4}$$

where $X_j \in \mathbb{R}^m$ is the expression data vector for gene j. It follows that the maximum-likelihood parameter values $\hat{\beta}_{ij}$ are the ordinary least-squares linear regression coefficients, $\hat{\omega}_j^2 = \frac{1}{m}\|X_j - \sum_{i\in \mathrm{Pa}_j} \hat{\beta}_{ij} X_i\|^2$ are the residual variances, and $\mathscr{L}_X$ evaluated at these maximum-likelihood values is the log of the total unexplained variance, up to an additive constant

$$\mathscr{L}_X = -\frac{m}{2}\sum_{j=1}^{n} \log(\hat{\omega}_j^2). \tag{S5}$$

- Adding more explanatory variables always reduces the residual variance in linear regression. Hence $\mathscr{L}_X$ as a function of $\mathscr{G}$ is maximized for maximal or fully connected DAGs with $n(n-1)/2$ edges. Such DAGs define a node ordering or permutation that turns $\mathbf{B}$ into a lower triangular matrix. Hence eq. (S5) can also be seen as a function on node orderings or permutations, and the maximum-likelihood values are then found by linearly regressing each node on its predecessors (i.e. parents) in the ordering [cf. eq. (S4)]. More precisely, let π be a permutation of $\{1,\ldots,n\}$, then
$$\mathscr{L}_{X,\pi} = \sum_{j=1}^{n}\left[-\frac{m}{2}\log(\omega_j^2) - \frac{1}{2\omega_j^2}\left\|X_j - \sum_{\pi_i<\pi_j}\beta_{ij}X_i\right\|^2\right] \tag{S6}$$
- Conversely, eq. (S3), and hence also eq. (S5), is easily seen to be invariant under any reordering of the nodes. Hence no edge directions can be inferred unambiguously from observational expression data without further constraints or information.

S1.2 Pairwise node ordering

To infer Bayesian gene networks, we first consider the log-likelihood score (6) without sparsity constraints,

$$\mathscr{L} \equiv \log P(\mathscr{G} \mid \mathbf{X},\mathbf{E}) = \log p(\mathbf{X} \mid \mathscr{G}) + \sum_{(i,j)\in\mathscr{G}} g_{ij}$$

where it is implicitly understood that the maximum-likelihood parameters are used in $\mathscr{L}_X = \log p(\mathbf{X} \mid \mathscr{G})$. Because $\mathscr{L}_X$ and $\mathscr{L}_P = \sum_{(i,j)\in\mathscr{G}} g_{ij}$ are both maximized for maximal DAGs, and because the value of $\mathscr{L}_X$ is the same for all maximal DAGs, it follows that to maximize $\mathscr{L}$, we need to find the maximal DAG or node ordering which maximizes the pairwise score $\mathscr{L}_P$. As stated in the main text, this is an NP-hard problem with no known polynomial approximation algorithms with a strong guaranteed error bound. The greedy algorithm we used is the standard heuristic for this type of problem [46].

S1.3 Sparsity constraints

Maximal DAGs lead to overfitting of the expression-based score $\mathscr{L}_X$, particularly in the case where the number of genes n is greater than the number of samples m. The most popular methods for imposing sparsity in Bayesian networks are:

- **Bayesian or Akaike Information Criterion.** The BIC or AIC methods augment the likelihood function $\mathscr{L}_X$ with a term proportional to the number of parameters in the model, i.e. the number of edges $|\mathscr{G}|$ in $\mathscr{G}$ (BIC $= -|\mathscr{G}|\log m$, AIC $= -|\mathscr{G}|$).

- **L1-penalized lasso regression.** In this case, the likelihood $\mathscr{L}_{X,\pi}$ [eq. (S6) is augmented by a term $\sum_{j=1}^{n} \lambda_j \sum_{\pi_i<\pi_j} |\beta_{ij}|$, such that finding the maximum-likelihood parameters $\hat{\beta}_{ij}$ becomes equivalent to performing a series of independent lasso regressions, one for each node on its predecessors in the ordering π. The extra penalty term can be understood as coming from a double-exponential prior distribution on the parameters β_{ij}.

An under-appreciated drawback of the BIC/AIC in high-dimensional settings is the fact that with a sufficient number of predictors it is possible to reduce ω_j^2 to zero for any gene, and hence make $\mathscr{L}_X$ (S5) arbitrarily large. By concentrating all interactions on one or a few target genes, this can be achieved while still keeping the BIC/AIC small. Hence in high-dimensional settings, use of the BIC/AIC leads to highly skewed 'all-or-nothing' in-degree distributions, as shown in Figure 2C, unless the maximum allowed number of regulators for each gene is capped at an artificially small number.

Similar problems can occur if lasso regression is used with a fixed λ for all j, because the number of candidate regulators differs greatly among genes that come early or late in the ordering. In [28], a method was proposed where the value of λ_j increased with the order of j, but their scaling could not provide any guarantee for the probability of false positive errors for individual edges in the resultant sparse graph. We used the lassopv variable selection method [50] instead. In brief, for each gene j and for each candidate regulator i of j (i.e. predecessor of j in the ordering π):

- calculate the largest (most stringent) value of λ_j for which i would be selected as a parent of j (i.e. have non-zero lasso regression coefficient);
- calculate the probability (p-value) of a randomly generated predictor having the same or larger 'critical' λ_j.

This results in a set of p-values p_{ij} for all pairs $\pi_i < \pi_j$, which achieve optimal false discovery control, i.e. they can be transformed into q-values q_{ij} by standard FDR correction methods such that if we keep all $q_{ij} \leq \alpha$, the expected FDR is less than α. Moreover for sufficiently small thresholds α, there is a corresponding penalty parameter value $\lambda_j(\alpha)$ such that the set of regulators with p_{ij} (or q_{ij}) less than α is precisely the set of regulators with non-zero lasso regression coefficient [50]. Hence in our method we can use thresholding on the p_{ij} directly to obtain sparse Bayesian networks.

In addition to the lasso regression based method for inducing sparsity, we also considered a simple **thresholding on the pairwise prior information** to obtain a sparse DAG. In eq. (6), if we set

$$g'_{ij} = \begin{cases} g_{ij} & \text{if } g_{ij} >= \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

then edges with $g_{ij} < \varepsilon$ are automatically excluded from the maximum-likelihood DAG, and the pairwise node ordering procedure will automatically result in a sparse DAG. This method does not provide any guarantee for the false positive control of individual edges in the (multi-variate) Bayesian network beyond what is provided by the pairwise causal inference test used.

S2 Supplementary figures

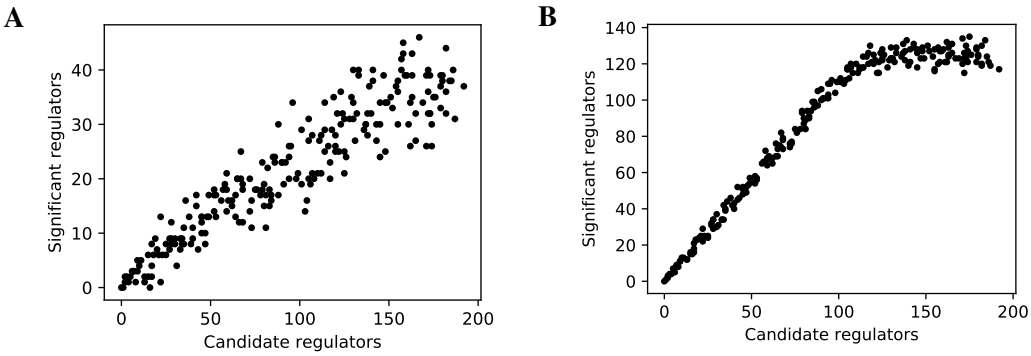

Figure S1: The linearity test of lasso-findr Bayesian networks at 5,000 (**A**) and 20,000 (**B**) significant interactions on DREAM dataset 1.

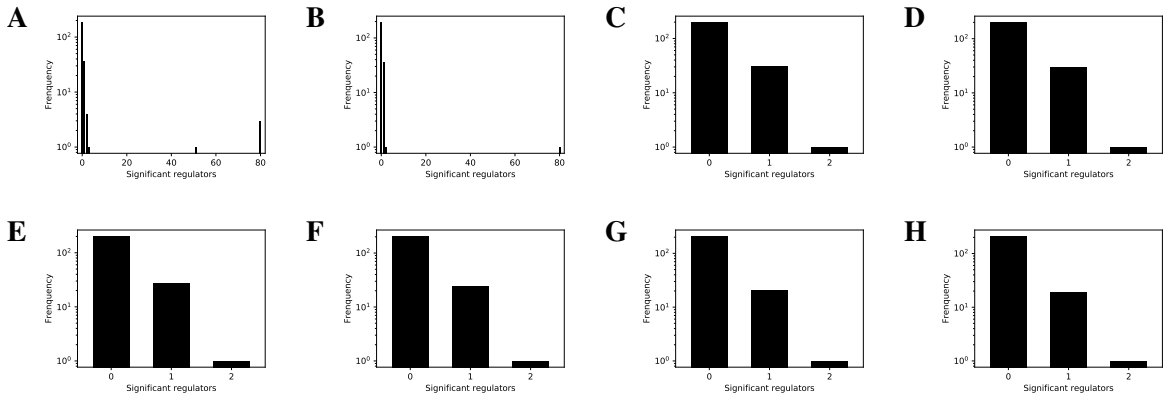

Figure S2: The histogram of significant regulator counts for each target gene in the bnlearn-hc Bayesian networks with AIC penalty 8.5 to 12 (**A** to **H**) and step 0.5 on DREAM dataset 1.

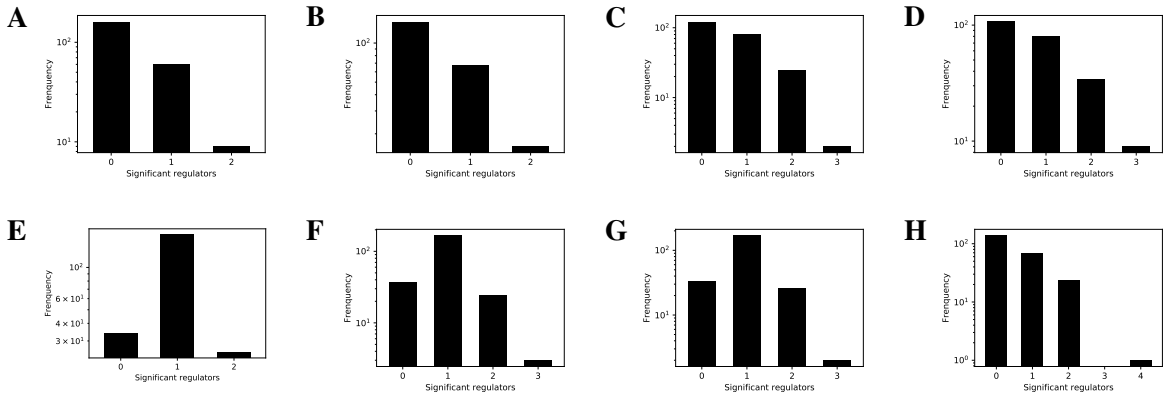

Figure S3: The histogram of significant regulator counts for each target gene in the bnlearn-fi Bayesian networks with nominal type I error rates 0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.2 (**A** to **H**) on DREAM dataset 1.

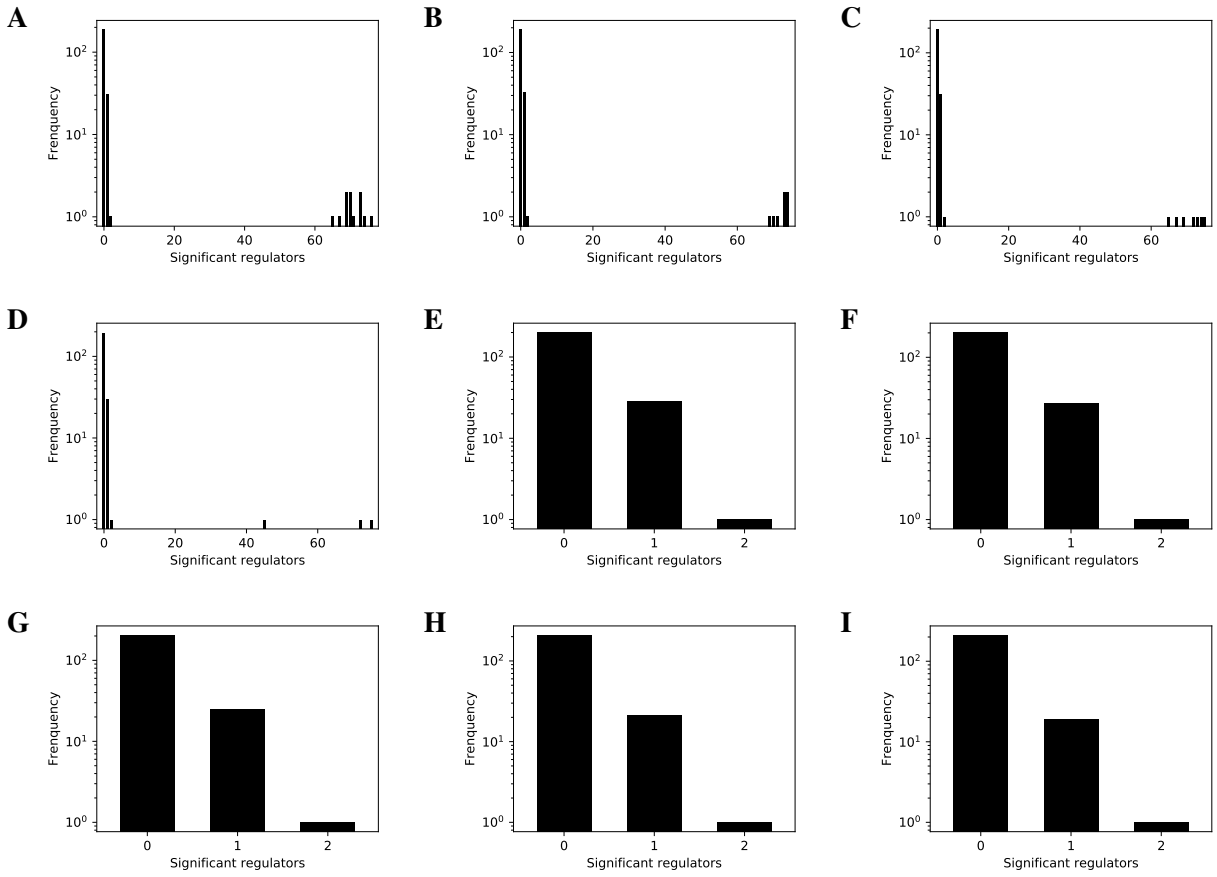

Figure S4: The histogram of significant regulator counts for each target gene in the bnlearn-hc-g Bayesian networks with AIC penalty 9.5 to 13 (**A** to **I**) and step 0.5 on DREAM dataset 1.

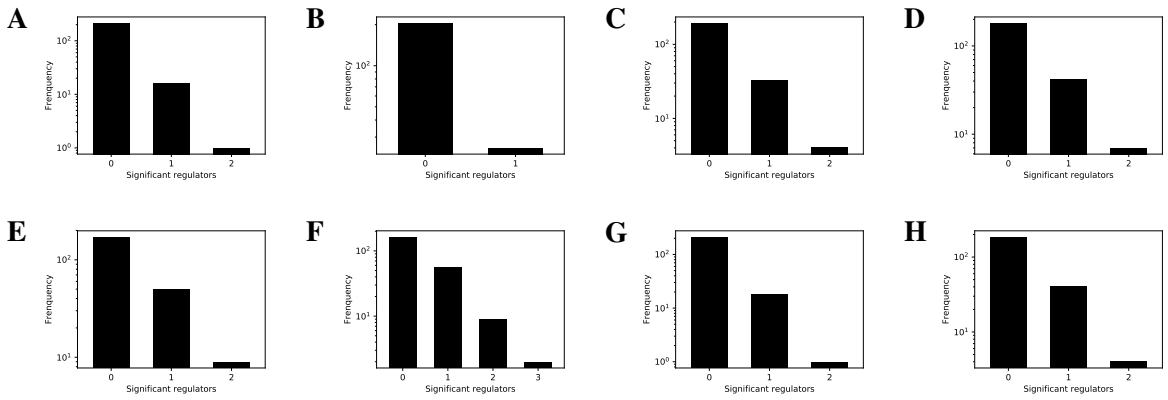

Figure S5: The histogram of significant regulator counts for each target gene in the bnlearn-fi-g Bayesian networks with nominal type I error rates 0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.2 (**A** to **H**) on DREAM dataset 1.

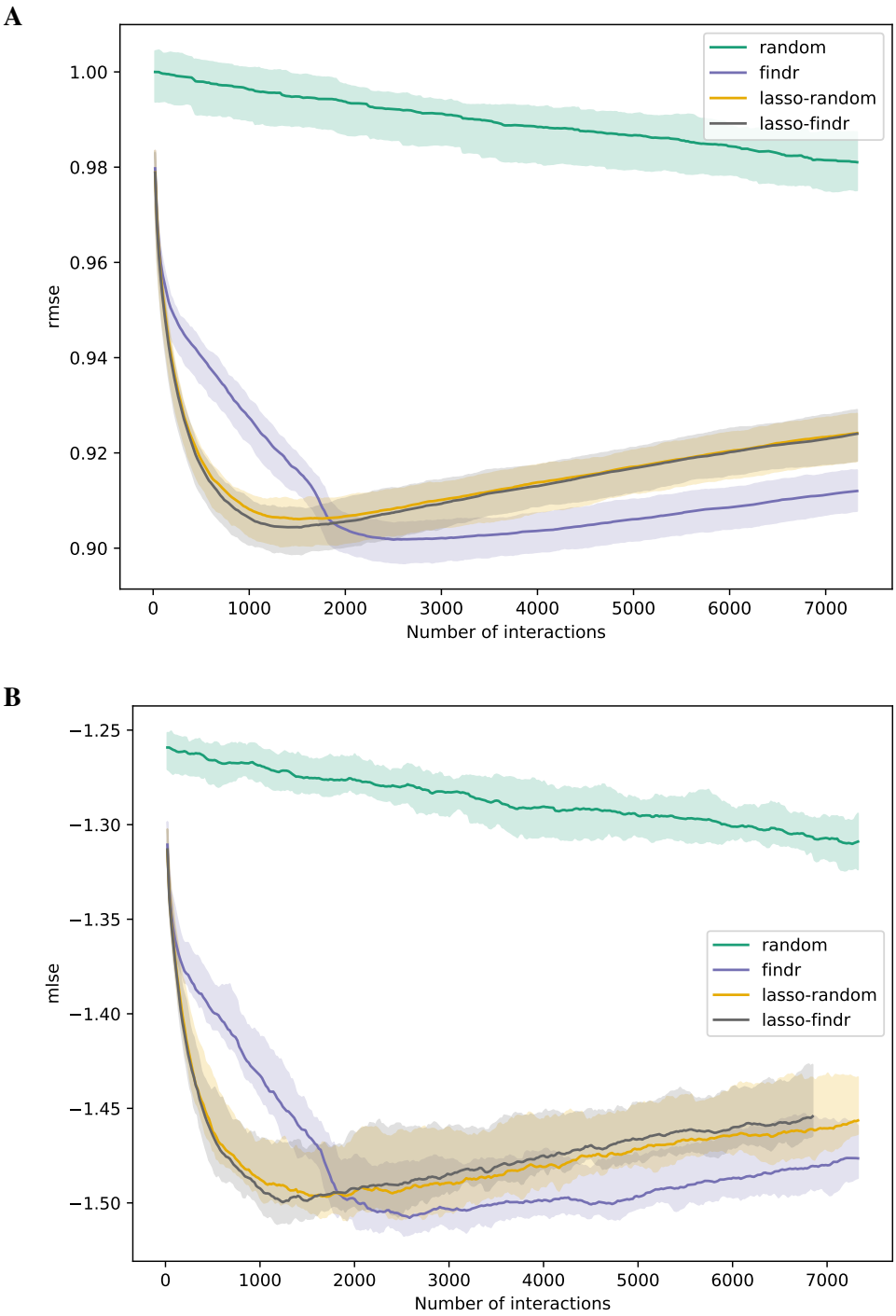

Figure S6: The root mean squared error (rmse, **A**) and mean log squared error (mlse, **B**) in training data are shown as functions of the numbers of predicted interactions in five-fold cross validations using linear regression models. Shades and lines indicate minimum/maximum values and means respectively. DREAM dataset 1 with 999 samples was used.

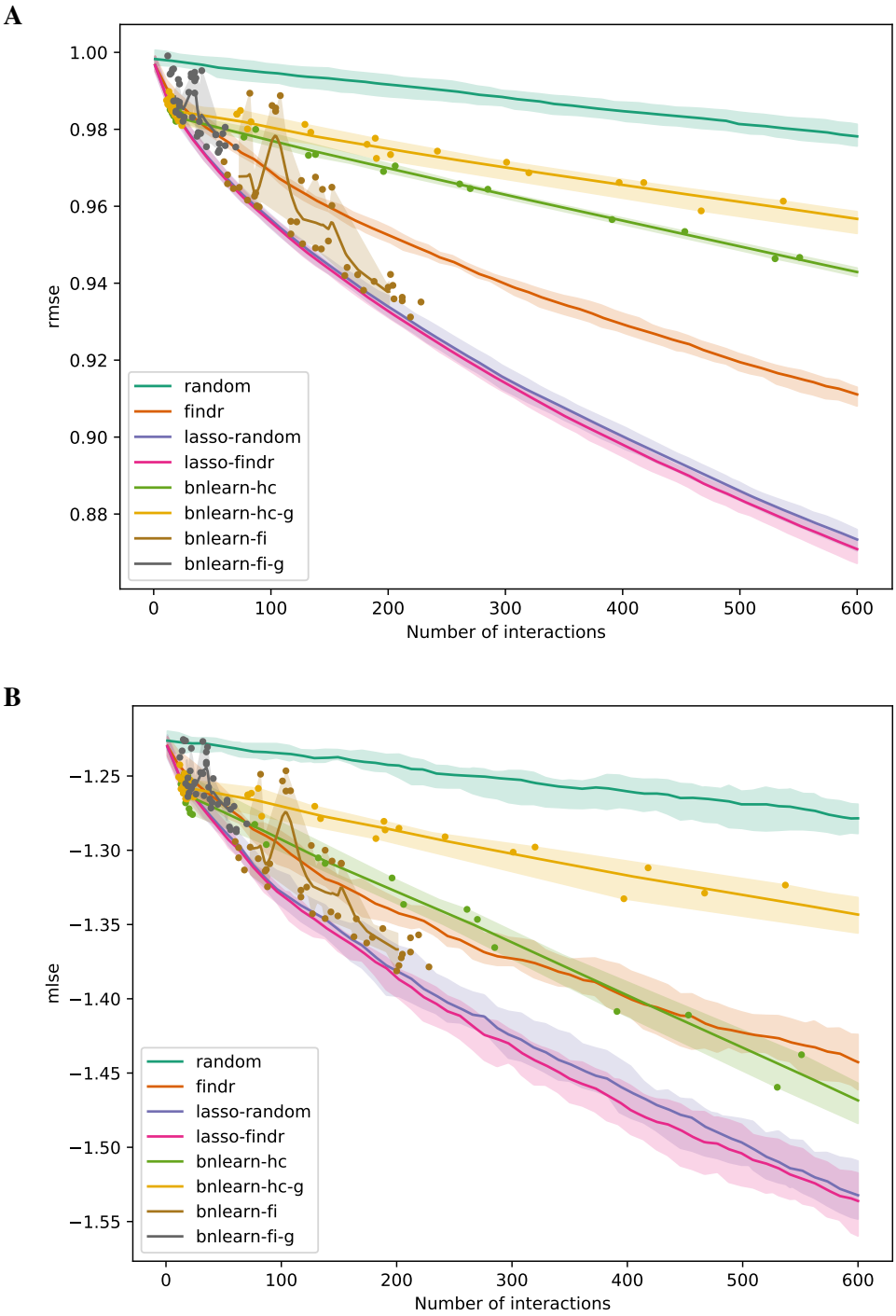

Figure S7: The root mean squared error (rmse, **A**) and mean log squared error (mlse, **B**) in training data are shown as functions of the numbers of predicted interactions in five-fold cross validations using linear regression models. Shades and lines indicate minimum/maximum values and means respectively. Root mean squared errors greater than 1 indicate over-fitting. DREAM dataset 1 with 100 samples was used.

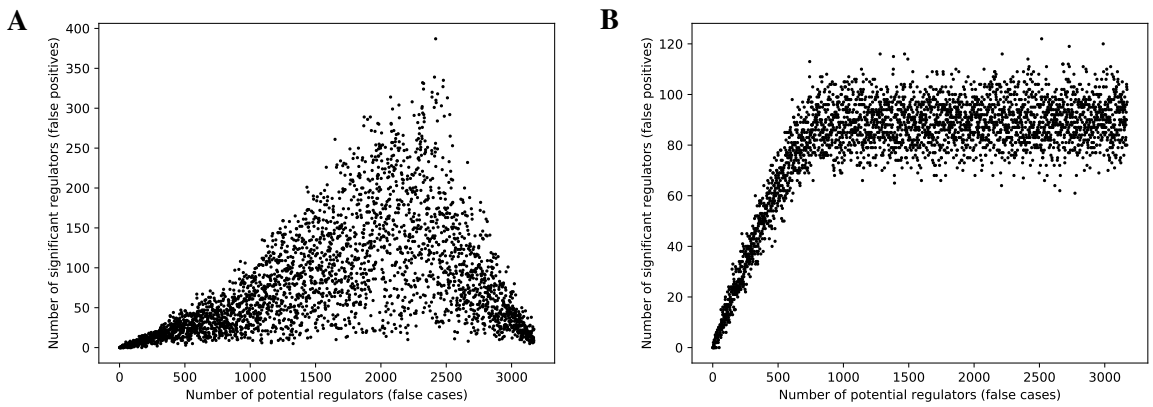

Figure S8: Conversion to Bayesian network from findr's predictions breaks its false discovery control.