High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 December 2019
doi: 10.3389/fgene.2019.01196

High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering

Lingfei Wang[1,2,3], *Pieter Audenaert*[4,5] *and Tom Michoel*[1,6]*

[1] Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush Campus, Midlothian, United Kingdom, [2] Broad Institute of Harvard and MIT, Cambridge, MA, United States, [3] Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, United States, [4] IDLab, Ghent University—imec, Ghent, Belgium, [5] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium, [6] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

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 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 conventional Bayesian network inference 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 node ordering by conducting pairwise causal inference tests between genes, which then allows to infer a Bayesian network *via* a series of independent variable selection problems, 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 conventional 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.

Keywords: systems genetics, network inference, Bayesian network, expression quantitative trait loci analysis, gene expression

OPEN ACCESS

Edited by:
Shizhong Han, Johns Hopkins Medicine, United States

Reviewed by:
Bochao Jia, Eli Lilly, United States
Sungwon Jung, Gachon University, South Korea

***Correspondence:**
Tom Michoel
tom.michoel@uib.no

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *03 July 2019*
Accepted: *29 October 2019*
Published: *20 December 2019*

Citation:
Wang L, Audenaert P and Michoel T (2019) High-Dimensional Bayesian Network Inference From Systems Genetics Data Using Genetic Node Ordering. Front. Genet. 10:1196. doi: 10.3389/fgene.2019.01196

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 (Rockman, 2008; Schadt, 2009; Civelek and Lusis, 2014; Albert and Kruglyak, 2015; Boyle et al., 2017). While important progress has been made in the experimental mapping of protein–protein and protein–DNA interactions (Walhout, 2006; Gerstein et al., 2012; Luck et al., 2017), the context-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 (Cusanovich et al., 2014). 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 traits of interest (Rockman, 2008; Schadt, 2009; Civelek and Lusis, 2014).

Gene network inference is a deeply studied problem in computational biology (Friedman, 2004; Albert, 2007; Bansal et al., 2007; Penfold and Wild, 2011; Emmert-Streib et al., 2012; Marbach et al., 2012; Äijö and Bonneau, 2016; Kiani et al., 2016). Among the many successful methods that have been devised, Bayesian networks are a powerful approach for modelling causal relationships and incorporating prior knowledge (Friedman et al., 2000; Friedman, 2004; Werhli and Husmeier, 2007; Mukherjee and Speed, 2008; Koller and Friedman, 2009; Pearl, 2009). In the context of complex trait genetics, the availability of genotype data leads to an especially significant prior on the direction of causality between correlated traits, which is based on the principle that genetic variation causes variation in gene expression or disease traits, but not *vice versa* (Schadt et al., 2005). Hence, Bayesian networks have become particularly popular for modelling conditional independence and causal dependence relationships among heritable traits, including molecular abundance traits (Zhu et al., 2004; Zhu et al., 2008; Neto et al., 2010; Hageman et al., 2011; Scutari et al., 2014). Using expression quantitative trait loci (eQTL) and gene expression data as input, Bayesian networks have been used for instance to identify key driver genes of type 1 diabetes (Schadt et al., 2008), Alzheimer's disease (Zhang et al., 2013; Beckmann et al., 2018), temporal lobe epilepsy (Johnson et al., 2015), and cardiovascular disease (Talukdar et al., 2016). However, Bayesian network inference is computationally demanding and limited to relatively small-scale systems. In this paper, we address the question whether Bayesian network inference from eQTL and gene expression data is feasible on a truely transcriptome-wide scale without sacrificing performance in terms of model fit and overlap with known interactions.

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 of a Bayesian network is usually inferred from the data using score-based or constraint-based approaches (Koller and Friedman, 2009). Score-based approaches maximize the likelihood of the model, or sample from the posterior distribution using Markov chain Monte Carlo (MCMC), using edge additions, deletions or inversions to search the space of network structures. Score-based methods have been shown to perform well using simulated genetics and genomics data (Zhu et al., 2007; Tasaki et al., 2015). Constraint-based approaches first learn the undirected skeleton of the network using repeated conditional independence tests, and then assign edge directions by resolving directional constraints (v-structures and acyclicity) on the skeleton. They have been used for instance in the joint genetic mapping of multiple complex traits (Scutari et al., 2014). However, the computational cost of both approaches is high. Because the number of possible graphs scales super-exponentially with the number of nodes, Bayesian gene network inference with conventional methods is feasible for systems of at most a few hundred genes or traits, and usually requires a hard limit on the number of regulators a gene can have as well as a preliminary dimension reduction step, such as filtering or clustering genes based on their expression profiles (Zhu et al., 2008; Zhang et al., 2013; Talukdar et al., 2016; Beckmann et al., 2018).

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 (Lappalainen et al., 2013; Franzén et al., 2016; GTEx Consortium, 2017). 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 (Boyle et al., 2017). The limitations of current Bayesian gene network inference methods mean that this model can be neither tested nor accomodated. Existing Bayesian network inference methods on categorical variables, e.g., Banjo (Smith et al., 2006), lack the resolution and directionality for transcriptomic datasets. 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 independent variable or feature selection problems, one for each node (Koller and Friedman, 2009; Shojaie and Michailidis, 2010). 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 (DAG) 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 (Bussemaker et al., 2007), or known protein-DNA and protein-protein interactions (Ernst et al., 2008; Greenfield et al., 2013), and those have been used together with score-based MCMC methods in Bayesian network inference previously (Werhli and Husmeier, 2007; Mukherjee and Speed, 2008).

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 (Schadt et al., 2005; Chen et al., 2007; Millstein

et al., 2009; Li et al., 2010; Neto et al., 2013; Millstein et al., 2016; Wang and Michoel, 2017a). 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 greedy optimization of the posterior log-likelihood of this generative model.

METHODS

An Algorithm for the Inference of Gene Regulatory Networks From Systems Genetics Data

To allow the inference of gene regulatory networks from high-dimensional systems genetics data, we developed a method that exploits recent algorithmic developments for highly efficient mapping of 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 (Shabalin, 2012; Qi et al., 2014; Ongen et al., 2015; Delaneau et al., 2017). 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 (Millstein et al., 2016) and Findr (Wang and Michoel, 2017a). In our experience, only Findr is sufficiently efficient to test for causality between millions of gene pairs.

C. Genetic Node Ordering

In *Bayesian Network Model for Systems Genetics Data*, 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 maximum-weight DAG, which induces a topological 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 maximum-weight DAG (**Figure 1D**). In this paper, we consider two approaches: *(i)* a truncation of the pairwise interaction scores retaining only the most confident (highest weight) edges in the maximum-weight DAG, and *(ii)* a multi-variate, L1-penalized lasso regression (Tibshirani, 1996; Wang and Michoel, 2017b) 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.

Bayesian Network Model With Prior Edge Information

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

$$p\left(x_1,...,x_n \mid G\right)=\prod_{j=1}^{n} p\left(x_j \mid \left\{x_i : i \in \mathrm{Pa}_j\right\}\right), \tag{1}$$

where Pa_j denotes the set of parent nodes of node j in the graph G. We only consider linear Gaussian networks (Koller and Friedman, 2009), 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 G is computed as

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

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

$$P\left(G \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid G\right)P\left(G\right),$$

where $P(G)$ is the prior probability of observing 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

$$\log P(G) \propto \sum_j \sum_{i \in \mathrm{Pa}_j} f_{ij},$$

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 G.

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 (Schadt et al., 2005; Chen et al., 2007; Millstein et al., 2009; Li et al., 2010; Neto et al., 2013; Millstein et al., 2016; Wang and Michoel, 2017a). Previously, such pairwise probabilities have been used as priors in conventional score-based Bayesian network inference (Zhu et al., 2004; Zhu et al., 2007), 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 (Neto et al., 2010; Hageman et al., 2011), 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 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 G can be written, upto a normalization constant, as

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

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

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

where E_i, $X_i \in \mathbb{R}^m$ are the ith rows of $\mathbf{E}$ and $\mathbf{X}$, respectively. $P(L_i \to G_i \to G_j | E_i, X_i, X_j)$ 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, and can be computed with pairwise causal inference methods (Millstein et al., 2016; Wang and Michoel, 2017a). In other words, conditional on a gene-to-gene DAG 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 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

$$P(\mathbf{E} \mid \mathbf{X},G) \propto \prod_j \prod_{i \in \mathrm{Pa}_j} e^{g_{ij}}, \tag{5}$$

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 $P(G) = \text{const}$, leads to an expression of the posterior log-likelihood that is formally identical to the model with prior edge information,

$$\log P(G \mid \mathbf{X},\mathbf{E}) = \log p(\mathbf{X} \mid G) + \sum_j \sum_{i \in \mathrm{Pa}_j} g_{ij} + \text{const.} \tag{6}$$

As before, if $g_{ij} = 0$, the edge (i,j) is excluded from being part of 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 $P(G) = \Sigma_j \Sigma_{i \in \mathrm{Pa}j} f_{ij}$, can still be added to the model, when such information is available.

Bayesian Network Parameter Inference

Given a DAG 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.

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 DAG with $n(n-1)/2$ edges. Such a DAG G defines a node ordering $\prec$ where $i \prec j \Leftrightarrow i \in \mathrm{Pa}_j$. 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 (Koller and Friedman, 2009) and **Supplementary Information**). Hence to maximize eq. (6) we need to find the node ordering or DAG which maximizes the term $\Sigma_j \Sigma_{i \in \mathrm{Pa}j} g_{ij}$. Finding the maximum-weight DAG is an NP-hard problem with no known polynomial approximation algorithms with a strong guaranteed error bound (Korte and Hausmann, 1978; Hassin and Rubinstein, 1994). 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 maximum-weight DAG

with $n(n-1)/2$ edges is obtained. Other edges are assigned probability 0 to indicate exclusion. The heuristic maximum-weight DAG reconstruction was implemented in Findr (Wang and Michoel, 2017a) as the command netr_one_greedy, with the *vertex-guided* algorithm for cycle detection (Haeupler et al., 2012).

Causal Inference of Pairwise Gene Regulations

We used Findr 1.0.6 (pij_gassist function) (Wang and Michoel, 2017a) 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.

Findr and Random Bayesian Networks From Node Orderings

The node ordering reconstruction removes less probable, cyclic edges, and results in a (heuristic) maximum-weight DAG G with edge weights $P_{ij} = e^{g_{ij}}$. We term these weighted DAGs as *findr* or *random Bayesian networks*, depending on the pairwise information used. A significance threshold can be applied on the continuous networks, to convert them to binary Bayesian networks at any desired sparsity level and thereby perform variable selection for the parents of every gene.

Lasso-Findr and Lasso-Random Bayesian Networks Using Penalized Regression on Ordered Nodes

As a second approach to perform variable selection in the maximum-weight DAGs, we performed hypothesis testing for every gene on whether each of its predecessors (in the findr or random Bayesian network) is a regulator, using L1-penalized lasso regression (Tibshirani, 1996) with the lassopv package (Wang and Michoel, 2017b) (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*.

Score-Based Bnlearn-Hc and Constraint-Based Bnlearn-Fi Bayesian Networks From Package Bnlearn

For comparison with score-based Bayesian network inference methods, we applied the hc function of the R package bnlearn (Scutari, 2010), 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 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 constraint-based Bayesian network inference methods [e.g., (Kalish and Buhlmann, 2007)], we applied the fast.iamb function of the R package bnlearn (Scutari, 2010), 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 (Neto et al., 2010; Hageman et al., 2011; Tasaki et al., 2015). Cis-eQTLs were only allowed to have outgoing 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.

Evaluation of False Discovery Control in Network Inference

Scoring metrics are comparable within each hypothesis test, but not neccesarily so between different hypothesis tests. Unlike p-values, the use of arbitray scores in network inference may lead to inconsistent false positive rates of candidate regulators among different target genes, which prevents consistent network-wide false discovery control (FDC) (Wang and Michoel, 2017b). However, the network-wide FDC consistency can be evaluated with the linear relation between the numbers of false positive regulators and candidate regulators for each gene. Violation of the linearity disproves the score for FDC in network inference. Due to the (in-degree) sparsity of biological networks, we discarded the top 5% of predictions to remove true positives, after which the FDC consistency was empirically evaluated with the linear relation between the numbers of false positive and candidate regulators. See (Wang and Michoel, 2017b) for method details.

Precision-Recall Curves and Points

We compared reconstructed Bayesian networks with gold standards using precision-recall (PR) 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.

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 (**Supplementary Information**, **Algorithm S1**). 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.

Data and Software

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 (Qi et al., 2014) to identify them, allowing for one false discovery per dataset. The DREAM challenge further provides the groundtruth network for each dataset, varying from around 1,000 to 5,000 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 (Lappalainen et al., 2013). 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: (Rockman, 2008) differential expression data from siRNA silencing experiments of transcription-associated factors (TFs) in a lymphoblastoid cell line (GM12878) (Cusanovich et al., 2014); (Schadt, 2009) DNA-binding information of TFs in the same cell line (Cusanovich et al., 2014); (Civelek and Lusis, 2014) the filtered proximal TF-target network from (Gerstein et al., 2012). 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. Processed Geuvadis data and groundtruth networks are available at https://github.com/lingfeiwang/findr-data-geuvadis

We preprocessed all expression data by converting them to a standard normal distribution separately for each gene, as explained in (Wang and Michoel, 2017a).

Software to reproduce the results from this study is available at the following URLs:

- Findr: https://github.com/lingfeiwang/findr.
- lassopv: https://github.com/lingfeiwang/lassopv.

RESULTS

Genetic Node Ordering Permits High-Dimensional Bayesian Network Inference

We developed a method for Bayesian network inference from high-dimensional systems genetics data which reconstructs a maximum-weight DAG from the confidence scores of pairwise causal inferences between gene expression traits using eQTLs as causal anchors, and which uses the node ordering induced by this DAG (termed "genetic node ordering" in reference to the use of genotype data to orient network edges) to decompose the Bayesian network inference task into a series of independent variable selection problems (*Methods*, *An Algorithm for the Inference of Gene Regulatory Networks From Systems Genetics Data*, **Figure 1**). Using an efficient implementation for the causal inference step (Wang and Michoel, 2017a), this approach allows to reconstruct Bayesian networks with thousands to ten-thousands of nodes. Our method is based on score-based Bayesian network inference methods for systems with pre-defined node orderings (Koller and Friedman, 2009; Shojaie and Michailidis, 2010), but differs in that the ordering is inferred from the same expression data, augmented with matched genotype data from the same samples, that is used for the subsequent Bayesian network log-likelihood maximization, using a single generative model (*Methods*, *Bayesian Network Model for Systems Genetics Data*), rather than relying on external prior information to determine the node ordering. Its computational efficiency is due to restricting the graph structure search space to Bayesian gene networks compatible with this inferred node ordering. This differs substantially from conventional score-based and constraint-based methods, including those that use genotype and gene expression data (Neto et al., 2010; Hageman et al., 2011; Tasaki et al., 2015), where the search space can only be reduced by limiting the possible number of parents for each gene to an artificially small number (Koller and Friedman, 2009). For clarity, a comparison of the main characteristics of the Bayesian network inference approaches considered in this paper is included in **Supplementary Table S1**.

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 (Storey and Tibshirani, 2003) which allows consistent, network-wide FDC (Chen et al., 2007; Wang and Michoel, 2017a). However, the enforcement of a gene ordering/Bayesian network partly broke the FDC, as the linearity between the numbers of false positive (i.e., significant here) and candidate regulators broke down at large candidate regulator counts (**Figure 2A**,

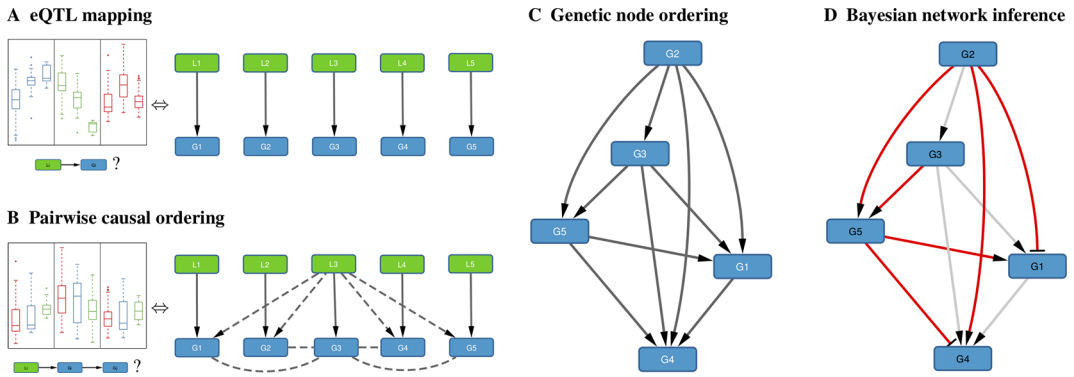

FIGURE 1 | Schematic overview of the method. **(A)** For each gene G_i, the *cis*-eQTL L_i whose genotype explains most of the variation in G_i expression is calculated; shown on the left are typical eQTL associations for three genes (colored blue, green, and red) where each box shows the distribution 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 $G_i \rightarrow G_j$. **(C)** A maximum-weight DAG having the genes as its nodes is derived from the pairwise causal interactions, which induces a "genetic" node ordering. **(D)** Variable selection is used to determine a sparse Bayesian gene network, which must be a sub-graph of the maximum-weight 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).

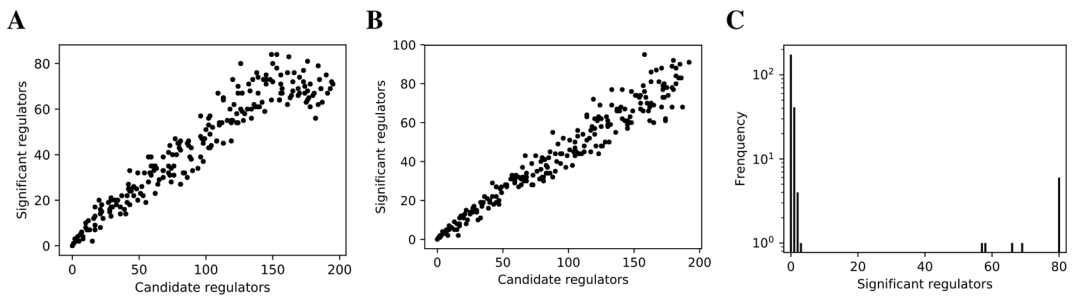

FIGURE 2 | False discovery controls of different Bayesian networks. **(A, B)** The linearity test of findr **(A)** and lasso-findr **(B)** Bayesian networks at 10,000 significant interactions on DREAM dataset 1. **(C)** The histogram of significant regulator counts for each target gene in the bnlearn-hc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

Methods). This effect is confirmed on the larger Geuvadis dataset in *Results on the Geuvadis Dataset Reaffirm Conclusions From Simulated Data*. By performing an extra lasso regression on top of the acyclic findr network, proper FDC was restored in terms of the linear relation in the lasso-findr Bayesian network (**Figure 2B**, **Supplementary Figure S1**).

In contrast, score-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 2C**, **Supplementary Figure S2**). This is in conflict with the known in-degree sparsity of gene regulation networks, which is required for its modularity, indicating that score-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.

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 score-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**).

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 PR curves, or points for the binary Bayesian networks from bnlearn-based methods, as shown in **Figure 3** with areas under the PR curve (AUPR) in **Supplementary Table S2**. Bnlearn based methods could only recover ∼2% of total true regulations, after which they suffered from a sharply dropping precision and behaved like random predictions. The highest precisions they achieved could not exceed those by lasso or findr based methods at the respective recalls either. In addition, bnlearn could not obtain >10% recall within 4-day time limit with any of the methods attempted. In this sense, 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 suboptimal than findr and lasso-based Bayesian networks.

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 five-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 rmse and mlse (the score optimized by bnlearn-hc). The findr Bayesian network explained the highest proportion of expression variation (≈2%) in the test data and identified the highest number of regulations (200 to 300), with runners up from lasso-based networks (≈1% variation, 50 regulations, **Figure 4**). The explained variance by findr and lasso networks grew to ≈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**).

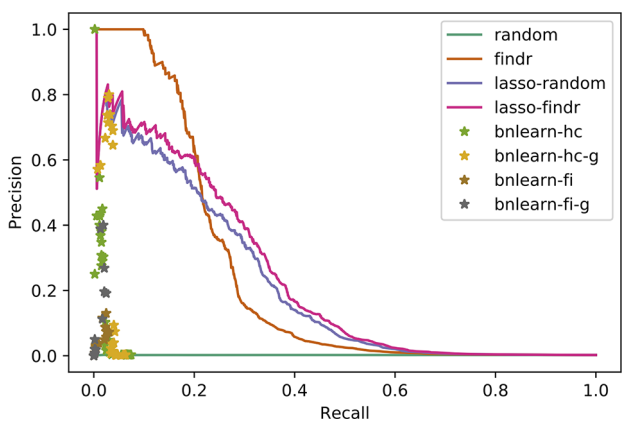

FIGURE 3 | Precision-recall curves/points of reconstructed Bayesian networks for DREAM dataset 1.

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 suggests that lasso may be capable of prioritizing edges with correct directions, and allows us to still recover genuine interactions even if the prior gene ordering is not fully accurate.

Lasso Bayesian Networks Mistake 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 6A**, and when also including genes without any cis-eQTL as only targets in **Figure 6B**. 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.

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

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*); conventional bnlearn-based network inference was attempted, but none of the restarts could complete within 1000 min.

Lasso-findr Bayesian networks were previously shown to provide ideal FDR control on this dataset (Wang and Michoel, 2017b), whereas findr Bayesian networks did not obtain a satisfying FDR control (**Supplementary Figure S8**). We believe this is due to the reconstruction of the node ordering, which interferes with the FDR control in pairwise causal inference. On the other hand, and again consistent with the DREAM data, findr

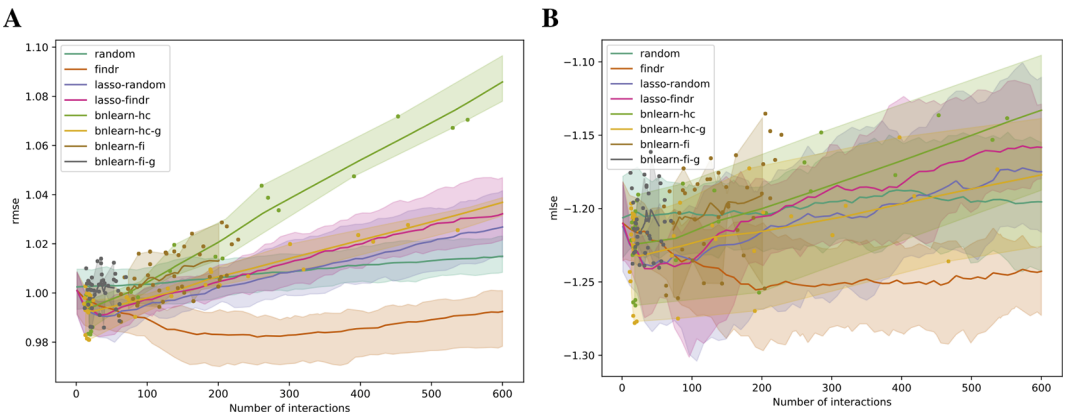

FIGURE 4 | The root mean squared error **(**rmse, **A)** and mean log squared error **(**mlse, **B)** in test data are shown as functions of the numbers of predicted interactions in five-fold cross validations using linear regression models. Shades and lines indicate minimum/maximum values and means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

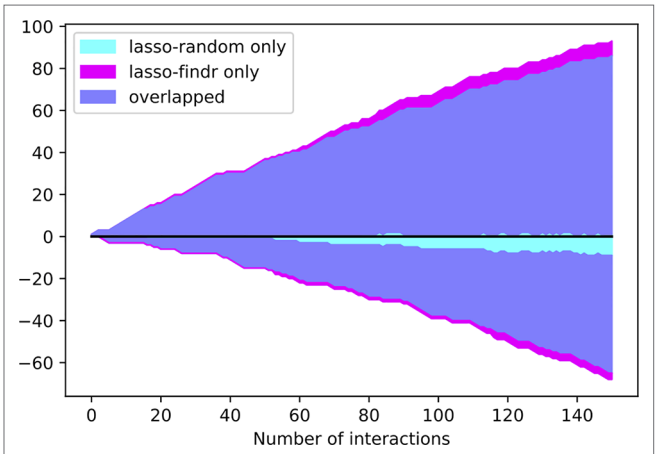

FIGURE 5 | The numbers of overlap and unique interactions (*y* axis) predicted by lasso-findr and lasso-random Bayesian networks as functions of the number of significant interactions in each network (*x* axis), on DREAM dataset 1. Positive and negative directions in *y* correspond to true and false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

Bayesian networks obtained superior results for the recovery of known transcriptional regulatory interactions inferred from ChIP-sequencing data (**Figures 7A**, **B**); neither method predicted TF targets inferred from siRNA silencing with high scores or accuracy better than random (**Figure 7C**).

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

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 conventional hill-climbing, MCMC sampling or constraint-based methods. Here we have introduced an alternative method, which first reconstructs a topological ordering of genes, and then infers a sparse maximum-likelihood Bayesian network using variable selection of parents for every gene from its predecessors in the ordering. Our method is applicable

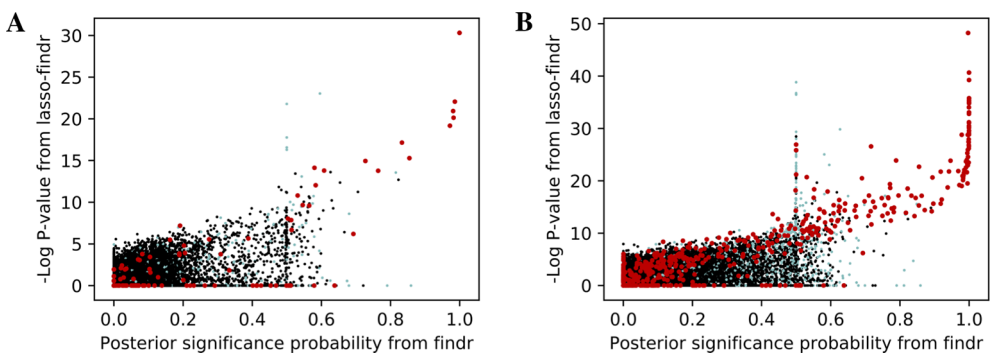

FIGURE 6 | The significance score of findr (posterior probability; x-axis) and in lasso-findr (-log P-value; y-axis) for direct true interactions (red), directly confounded gene pairs (cyan), and other, unrelated gene pairs (black) on DREAM dataset 1; in **(A)** only genes with cis-eQTLs are considered 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 | Timings for different Bayesian network inference methods/programs.

Dataset	Samples	Genes	Findr	lassopv	bnlearn-hc	bnlearn-fi
DREAM	100	230	< 1 s	≈1 min	≥10 h	≥30 min
Geuvadis	360	3172	< 1 min	≈10 h	–	–

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.

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 conventional score-based methods, irrespective of their computational cost, was their over-fitting 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 score-based 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 reconstructs a Bayesian network as a sparse subgraph from a maximum-weight DAG determined by pairwise causal relationships inferred using instrumental variable methods. We considered two variants of the method: one where the edge weights in the maximum-weight DAG 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 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 (Wang and Michoel, 2017b), 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 "naïve" 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 (Zou and Hastie, 2005), 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 FDC did not overcome the limitations introduced by confounding interactions.

On the other hand, the lasso-random network used solely transcriptomic profiles, yet provided better performance than the conventional score-based and constrained-based networks, including those that used genotypic information. Together with its better FDC, this makes the lasso-random network an interesting method for high-dimensional 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 (Koller and Friedman, 2009; Pearl, 2009), i.e., there is usually a large class of Bayesian networks with very different topology which all explain the

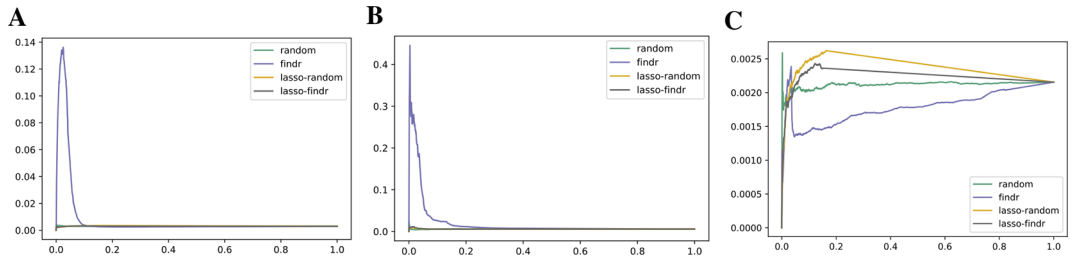

FIGURE 7 | Precision-recall curves for Bayesian networks reconstructed from the Geuvadis dataset for three groundtruth networks: DNA-binding of 20 TFs in GM12878 **(A)**, DNA-binding of 14 TFs in five ENCODE cell lines **(B)**, and siRNA silencing of six TFs in GM12878 **(C)**.

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. This is because human ground-truth networks have strong limitations. They are normally reconstructed from heterogeneous, noisy, high-throughput data (e.g., ChIP-sequencing and/or knock-out experiments), and are both incomplete (many true interactions are not present) and imperfect (many detected physical interactions have no functional effect). In addition, statistical inference algorithms can hardly distinguish direct interactions from indirect ones, which operate through an unidentified third factor and should be regarded as "false positives". 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 (not only transcription factors), 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 (Subramanian et al., 2017), 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 the conventional score-based algorithms, and hence a comparison would not be fair. 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 genetic node ordering is a highly efficient approach for reconstructing gene regulatory networks from high-dimensional systems genetics data, which outperforms conventional methods by restricting the super-exponential graph structure search space to acyclic graphs compatible with the causal inference results, and which is sufficiently flexible to integrate other types of pairwise prior data when they are available.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.synapse.org//#!Synapse:syn2820440/wiki/, https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/, https://github.com/lingfeiwang/findr-data-geuvadis. Findr: https://github.com/lingfeiwang/findr-R and https://github.com/lingfeiwang/findr, lassopv: https://github.com/lingfeiwang/lassopv, bnlearn: http://www.bnlearn.com/.

AUTHOR CONTRIBUTIONS

Conceptualization: LW, TM. Data curation: LW. Formal analysis: LW, PA, TM. Funding acquisition: TM. Investigation: LW, PA, TM. Methodology: LW, PA, TM. Software: LW. Supervision: TM. Writing: LW, PA, TM.

FUNDING

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

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (Wang et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01196/full#supplementary-material

REFERENCES

Äijö, T., and Bonneau, R. (2016). Biophysically motivated regulatory network inference: progress and prospects. *Hum. Heredity* 81 (2), 62–77. doi: 10.1159/000446614

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Albert, R. (2007). Network inference, analysis, and modeling in systems biology. *Plant Cell* 19 (11), 3327–3338. doi: 10.1105/tpc.107.054700

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78. doi: 10.1038/msb4100120

Beckmann, N. D., Lin, W. J., Wang, M., Cohain, A. T., Wang, P., Ma, W., et al. (2018). Multiscale causal network models of Alzheimer's disease identify VGF as a key regulator of disease. *bioRxiv* p, 458430. doi: 10.1101/458430

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell.* 169 (7), 1177–1186. doi: 10.1016/j.cell.2017.05.038

Bussemaker, H. J., Foat, B. C., and Ward, L. D. (2007). Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.* 36, 329–347. doi: 10.1146/annurev.biophys.36.040306.132725

Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* 8 (10), R219. doi: 10.1186/gb-2007-8-10-r219

Civelek, M., and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* 15 (1), 34–48. doi: 10.1038/nrg3575

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PloS Genet.* 10 (3), e1004226. doi: 10.1371/journal.pgen.1004226

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PloS Genet.* 10 (3), e1004226. doi: 10.1371/journal.pgen.1004226

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* 8, 15452. doi: 10.1038/ncomms15452

Emmert-Streib, F., Glazko, G., Altay, G., and De Matos Simoes, R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* 3, 8. doi: 10.3389/fgene.2012.00008

Ernst, J., Beg, Q. K., Kay, K. A., Bala'zsi, G., Oltvai, Z. N., and Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor - gene interactions in Escherichia coli. *PloS Comp. Biol.* 4, e1000044. doi: 10.1371/journal.pcbi.1000044

Franzén, O., Ermel, R., Cohain, A., Akers, N., Di Narzo, A., Talukdar, H., et al. (2016). Cardiometabolic risk loci share downstream cis and trans genes across tissues and diseases. *Science* 827–830. doi: 10.1126/science.aad6970

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* 308, 799–805. doi: 10.1126/science.1094068

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489 (7414), 91–100. doi: 10.1038/nature11245

Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics.* 29 (8), 1060–1067. doi: 10.1093/bioinformatics/btt099

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature* 550 (7675), 204. doi: 10.1038/nature24277

Haeupler, B., Kavitha, T., Mathew, R., Sen, S., and Tarjan, R. E. (2012). Incremental cycle detection, topological ordering, and strong component maintenance. *ACM Trans. Algorithms* 8 (1), 3:1–3:33. doi: 10.1145/2071379.2071382

Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., and Churchill, G. A. (2011). A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics* 187 (4), 1163–1170. doi: 10.1534/genetics.110.123273

Hassin, R., and Rubinstein, S. (1994). Approximations for the maximum acyclic subgraph problem. *Inf. Process. Lett.* 51 (3), 133–140. doi: 10.1016/0020-0190(94)00086-7

Johnson, M. R., Behmoaras, J., Bottolo, L., Krishnan, M. L., Pernhorst, K., Santoscoy, P. L. M., et al. (2015). Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. *Nat. Commun.* 6, 6031. doi: 10.1038/ncomms7031

Kalisch, M., and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC Algorithm. *J. Mach. Learn. Res.* 8, 613–636.

Kiani, N. A., Zenil, H., Olczak, J., and Tegnér, J. (2016). Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. *Semin. Cell Dev. Biol.* 51, 44–52. doi: 10.1016/j.semcdb.2016.01.012

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA, USA: The MIT Press).

Korte, B., and Hausmann, D. (1978). "An analysis of the greedy heuristic for independence systems," in *Annals of Discrete Mathematics*, vol. 2. 65–74. doi: 10.1016/S0167-5060(08)70322-4

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Lappalainen, T., Sammeth, M., Friedlander, M. R., Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 09501 (7468), 506–511. doi: 10.1038/nature12531

Li, Y., Tesson, B. M., Churchill, G. A., and Jansen, R. C. (2010). Critical reasoning on causal inference in genome wide linkage and association studies. *Trends Genet.* 26 (12), 493–498. doi: 10.1016/j.tig.2010.09.002

Luck, K., Sheynkman, G. M., Zhang, I., and Vidal, M. (2017). Proteome-scale human interactomics. *Trends Biochem. Sci.* 42 (5), 342–354. doi: 10.1016/j.tibs.2017.02.006

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9 (8), 796–804. doi: 10.1038/nmeth.2016

Millstein, J., Zhang, B., Zhu, J., and Schadt, E. E. (2009). Disentangling molecular relationships with a causal inference test. *BMC Genet.* 10 (1), 23. doi: 10.1186/1471-2156-10-23

Millstein, J., Chen, G. K., and Breton, C. V. (2016). cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics* 32, 2364–2365. doi: 10.1093/bioinformatics/btw135

Mukherjee, S., and Speed, T. P. (2008). Network inference using informative priors. *Proc. Natl. Acad. Sci.* 105 (38), 14313–14318. doi: 10.1073/pnas.0802272105

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann. Appl. Stat.* 4 (1), 320. doi: 10.1214/09-AOAS288

Neto, E. C., Broman, A. T., Keller, M. P., Attie, A. D., Zhang, B., Zhu, J., et al. (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* 193 (3), 1003–1013. doi: 10.1534/genetics.112.147124

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2015). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32 (10), 1479–1485. doi: 10.1093/bioinformatics/btv722

Pearl, J. (2009). *Causality* (Cambridge, UK: Cambridge University Press). doi: 10.1017/CBO9780511803161

Penfold, C. A., and Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus.* 1 (6), 857–870. doi: 10.1098/rsfs.2011.0053

Qi, J., Foroughi Asl, H., Björkegren, J. L. M., and Michoel, T. (2014). kruX: Matrix-based non-parametric eQTL discovery. *BMC Bioinf.* 15, 11. doi: 10.1186/1471-2105-15-11

Qi, J., Foroughi Asl, H., Bjorkegren, J., and Michoel, T. (2014). kruX: matrix-based non-parametric eQTL discovery. *BMC Bioinf.* 15 (1), 11. doi: 10.1186/1471-2105-15-11

Rockman, M. V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* 456 (7223), 738–744. doi: 10.1038/nature07633

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37 (7), 710–717. doi: 10.1038/ng1589

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PloS Biol.* 6, e107. doi: 10.1371/journal.pbio.0060107

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1038/nature08454

Scutari, M., Howell, P., Balding, D. J., and Mackay, I. (2014). Multiple quantitative trait analysis using Bayesian networks. *Genetics* 198 (1), 129–137. doi: 10.1534/genetics.114.165704

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Software* 35 (1), 1–22.

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28 (10), 1353–1358. doi: 10.1093/bioinformatics/bts163

Shojaie, A., and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs. *Biometrika.* 97 (3), 519–538. doi: 10.1093/biomet/asq038

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J., and Jarvis, E. D. (2006). Computational Inference of Neural Information Flow Networks. *PloS Comput. Biol.* 2 (11), e161. doi: 10.1371/journal.pcbi.0020161

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* Aug100 (16), 9440–9445. doi: 10.1073/pnas.1530509100

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171 (6), 1437–1452. doi: 10.1016/j.cell.2017.10.049

Talukdar, H., Foroughi Asl, H., Jain, R., Ermel, R., Ruusalepp, A., Franzén, O., et al. (2016). Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst.* 2, 196–208. doi: 10.1016/j.cels.2016.02.002

Tasaki, S., Sauerwine, B., Hoff, B., Toyoshiba, H., Gaiteri, C., and Neto, E. C. (2015). Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. *Genetics* 199 (4), 973–989. doi: 10.1534/genetics.114.172619

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. (Methodol.)* p, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Walhout, A. J. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* 16 (12), 1445–1454. doi: 10.1101/gr.5321506

Wang, L., and Michoel, T. (2017a). Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data. *PloS Comput. Biol.* 13 (8), e1005703. doi: 10.1371/journal.pcbi.1005703

Wang, L., and Michoel, T. (2017b). doi: 10.1101/288217 Controlling false discoveries in Bayesian gene networks with lasso regression p-values. arXiv:170107011 q-bio, stat.Jan;ArXiv: 1701.07011. Available from: http://arxiv.org/abs/1701.07011.

Wang, L., Audenaert, P., and Michoel, T. (2019). High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. *bioRxiv.* doi: 10.1101/501460

Werhli, A. V., and Husmeier, D. (2007). 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. doi: 10.2202/1544-6115.1282

Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* Apr153 (3), 707–720. doi: 10.1016/j.cell.2013.03.030

Zhu, J., Lum, P. Y., Lamb, J., GuhaThakurta, D., Edwards, S. W., Thieringer, R., et al. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* 105, 363–374. doi: 10.1159/000078209

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., et al. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PloS Comput. Biol.* 3 (4), e69. doi: 10.1371/journal.pcbi.0030069

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., et al. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40, 854–861. doi: 10.1038/ng.167

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B. (Stat. Methodol.)* 67 (2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.