Nonparanormal Distributions & Causal Inference with Single-Cell RNA-Seq Data

Elizabeth Silver

June 14, 2016

Abstract

Background. Single-cell RNA-Seq is a new technique that can measure gene expression levels in individual cells. We would like to use single-cell RNA-seq data to learn genetic regulatory networks. This is a natural task for causal-model structure-learning algorithms, which aim to learn the causal relationships between the measured variables. Causal algorithms perform poorly in high dimensions unless the data are Gaussian, and single-cell RNA-Seq data are non-Gaussian. However, the "nonparanormal SKEPTIC" method extends causal algorithms to high-dimensional Gaussian copula distributions, which may better approximate single-cell RNA-Seq data.
Aim. To learn a genetic regulatory network by applying the SKEPTIC to real single-cell gene expression data, validating against known regulatory interactions.
Data. 24,175 gene expression levels in 934 mouse embryonic stem cells were measured using inDrop single-cell RNA-seq. 500 high-variance genes, including 120 transcription factors, were selected for network recovery.
Method. The covariance matrix over the single-cell RNA-Seq data was estimated using the SKEPTIC, and input to causal algorithms, producing a graph over all measured genes. The performance was evaluated on (a) a set of known transcription factor binding relationships from ChIP-Seq studies, and (b) regulatory effects learned from loss-of-function/gain-of-function experiments.
Results. Previous studies did no better than chance at identifying adjacencies for eukaryotic organisms. Applying the SKEPTIC to single-cell data and using FGS for structure learning, we identified adjacencies with 22.5% precision, a $14\times$ improvement over chance ($p < 10^{-45}$).
Conclusion. Single-cell RNA-Seq data may be used for automatic, accurate recovery of the genetic regulatory network. These networks help to organize everything from embryonic development to cancer progression. Thus, these methods can be applied in both developmental genetics and personalized cancer medicine.

Keywords: Gaussian copula, causal graphical model, single-cell RNA-seq, mESCs

1 Introduction

Genetic regulatory networks govern cellular replication, differentiation, and much of cellular response to stimuli. They encode *causal* relationships between variables, not just predictive relationships. Causal relationships tell us about the consequences of *intervening* in the system. If Gene A regulates Gene B, then if we perform an experiment to knock out or overexpress Gene A, we should see a change in the expression of Gene B. Predictive information only licenses inferences about data drawn from the same distribution as our training data, whereas causal information allows us to make inferences about different distributions: distributions after intervention.

Experiments are the ideal way to learn causal relationships, but they are expensive. To learn an entire genetic regulatory network would require experimenting on all combinations of genes. This is infeasible: even *E. coli* has ∼4,300 genes. It would be helpful to have a method for learning causal relationships from observational data, or prioritizing the experiments most likely to be informative.

Learning causal relationships from observational data has been an active area of research in machine learning since the late 1980s (Spirtes et al., 1989, 2000; Pearl et al., 1991; Pearl, 2009). In the context of genetic regulation, there are many challenges to inferring causal relationships:

1. Cyclic causal structure (e.g. A may regulate B, and B regulate A)
2. Many unobserved confounding variables, both exogenous (e.g. environmental features) and endogenous (e.g. post-transcriptional regulation)
3. Small sample sizes
4. High-dimensional variable sets
5. Measurement noise
6. Aggregated data (microarrays can only measure the *average* expression level of each gene across thousands of cells)

Aggregated data is particularly problematic. Chu et al. (2003) argued that there is no reason to think that we can learn causal structure from microarray data using standard causal inference methods. Conditional independence relationships that hold in the distribution over individual cells are not guaranteed to hold in the distribution over aggregates of cells if the dependencies between variables are nonlinear (as they are for many genetic regulatory mechanisms). Conditional independences are crucial to inferring causal structure.

Happily, a new technique produces non-aggregated measurements. Single-cell RNA-seq uses quantitative seqencing technology to measure the number of mRNA transcripts of each gene within an individual cell (Klein et al., 2015). Measuring gene expression at the single cell level allows us to observe inter-cell variation that would otherwise be obscured (Wills et al., 2013). Also, measuring individual cells makes it relatively cheap to increase the sample size, an important bonus in this high-dimensional learning problem.

Algorithms for learning causal relationships between variables (henceforth "causal search algorithms") have been scaled to high-dimensional problems (Maathuis et al., 2009; Colombo and Maathuis, 2012) under the assumption of multivariate Gaussianity. However, single-cell RNA-seq data are non-Gaussian. They may be better modeled by Gaussian copula or "nonparanormal" distributions. Liu et al. (2012) have developed the SKEPTIC, an efficient method for learning nonparanormal distributions. Harris and Drton (2013) showed that the SKEPTIC allows us to learn causal relationships from nonparanormal data.

Evaluating causal network reconstruction is more difficult than evaluating predictive algorithms, because we are trying to learn the effects of performing *interventions*, which would change the distribution of the data. It is not possible to use within-sample risk to estimate out-of-sample risk. Because the true genetic regulatory network is unknown, we must use a "silver standard" – a subset of the true causal relationships, learned from experiments – and evaluate how well the algorithm recovers those relationships.

2 Problem and Approach

We aim to set a benchmark for genetic regulatory network construction from single-cell RNA-seq data. We will perform a rigorous evaluation against known regulatory relationships from ChIP-chip/seq and Loss Or Gain Of Function (LOGOF) experiments, and compare our results to those of related studies. We will also test whether using the nonparanormal SKEPTIC improves performance.

3 Background & Related Work

3.1 Causal modeling

We use directed graphs to represent causal relationships between variables – e.g. "A causally influences B" is represented as $A \to B$. Readers unfamiliar with the causal graphical modeling literature may refer to the Appendix, Section 13 for a short summary of key results and terminology. We use the following notation:

Notation and definitions. A directed graph $G = (V, E)$ consists of a vertex set, V, and an edge set E. The edges $\langle v_1, v_2 \rangle \in E$ are ordered pairs of vertices. We usually work with *directed acyclic graphs* (DAGs), which contain no directed cycles.

The vertices or "nodes" represent random variables. The set of nodes adjacent to a node v in G are called $neighbors(v)$. Of the neighbors, those with edges into v are called $parents(v)$, and those with edges out are called $children(v)$. If there is a directed path of any length $A \to \cdots \to B$, then A is an *ancestor* of B and B is a *descendant* of A. A *V-structure* or "unshielded collider" is a triple of nodes A, B, C such that A is a parent of C and B is a parent of C, but A and C are not adjacent.

In a *causal* graphical model, an edge $A \to B$ represents the statement that manipulating A will change the distribution of B, all else equal. That is to say, there are at least two values, $a \neq a'$, two distributions, $p(B) \neq p'(B)$, and some set of values of the variables $V \setminus \{A, B\} = v$, such that if one were to intervene on A and set its value to either a or a', while holding all other variables $V \setminus \{A, B\}$ constant at v, the distribution of B would be $p(B)$ or $p'(B)$, respectively.

In DAGs, the d-separation criterion (see Section 13.2) gives a complete list of the independences implied by the Causal Markov Assumption (see Section 13.1).[1] The set of DAGs that entail the same conditional independences as a DAG G is called the *Markov Equivalence Class* (MEC) of G. The MEC of a DAG can be represented by a *Complete Partially Directed Acyclic Graph* (CPDAG). If we marginalize some nodes in a causal DAG, we can represent the causal relationships among the remaining nodes using a *Maximal Ancestral Graph* (MAG). The MEC of a MAG can be represented by a *Partial Ancestral Graph* (PAG) (see Section 13.4).

Causal search algorithms. This study will compare four causal search algorithms: PC-stable, FGS, FCI, and GFCI. Table 1 summarizes the differences between them.

Assumes:	PC-stable	FGS	FCI	GFCI
Allows latent variables?	No; returns CPDAG	No; returns CPDAG	Yes; returns PAG	Yes; returns PAG
Particular distributions?	Nonparametric	Gaussian or multinomial	Nonparametric	Gaussian or multinomial
Faithfulness?	Yes	Weaker[2]	Yes	Yes

Table 1: Different assumptions made by four causal search algorithms

PC-stable (Colombo and Maathuis, 2012) is a variant of the PC[3] algorithm (Spirtes et al., 2000). PC has two phases: Fast Adjacency Search (FAS) and Orientation.

FAS begins with a complete undirected graph and performs a series of conditional independence tests. When nodes X_i and X_j are found conditionally independent given some set $\mathbf{S}$, the edge between X_i and X_j is removed. The trick is the order of tests: unconditional independences are tested first, then first-order conditional independences, etc. The sets $\mathbf{S}$ are chosen from subsets of $neighbors(X_i) \cup neighbors(X_j)$, because by the Causal Markov Condition (Section 13.1), a node's parents screen it off from its non-descendants. As edges are removed, the set of valid conditioning sets shrinks, reducing the number of future tests. This introduces order-dependence: the results of earlier tests determine the valid conditioning sets of later tests. Unlike

[1] ↑This is not true in general for *cyclic* directed graphs, except in the special case where the models are linear-Gaussian and the data are from the equilibrium distribution (Richardson, 1996).

[2] ↑FGS assumes that the graph with the highest score is correct. This is weaker than faithfulness but more restrictive than minimality.

[3] ↑'PC' just stands for 'Peter and Clark', after Peter Spirtes and Clark Glymour.

PC, PC-Stable only removes edges after all tests of a given conditioning set size have been completed, reducing the order-dependence.

Both PC and PC-stable then orient edges by identifying V-structures, and orienting all the other edges that cannot be flipped without creating additional V-structures. In the limit, both PC and PC-stable return the true CPDAG, but in practice PC-stable performs much better in high-dimensional problems. PC is nonparametric as it can be used with any conditional independence test. However, no efficient, accurate, high-dimensional nonparametric independence tests exist, so parametric tests are required for high dimensional problems.

FGS. Fast Greedy Search (FGS) is a variant of Greedy Equivalence Search (GES) (Chickering, 2002; Chickering and Meek, 2002). GES has two phases: forward and backward. The forward phase starts with an empty graph, and proceeds by adding whichever edge most improves the BIC score. After each addition, the Markov Equivalence Class of the model is computed, then the next highest-scoring edge is added. GES thus moves through a space of MECs, optimizing the BIC. Once the forward phase reaches an optimum, the backward phase starts removing edges. Some edges added in the forward phase may be redundant, and removing them improves the score because BIC penalizes complexity. Two phases suffice: in the limit, GES finds the correct CPDAG. FGS is a variant of GES that achieves dramatic increases in speed via some modifications to the data structures (Ramsey, 2015).

FCI. Fast Causal Inference (FCI) is a constraint-based search similar to PC, but relaxing the assumption that there are no latent variables (more precisely, the assumption of "Causal Sufficiency" – see Section 13.1) so it outputs a PAG instead of a CPDAG. Like PC, FCI begins with the Fast Adjacency Search, but has a more complex orientation phase. For full details regarding FCI, see Spirtes et al. (2000).

GFCI. Greedy Fast Causal Inference (GFCI) is a hybrid search combining FGS with FCI. It begins by running FGS, then uses FCI as a post-processor on the output of FGS. FCI performs tests which may cause it to remove some edges and reorient others, turning the CPDAG produced by FGS into a PAG. GFCI is a new algorithm; details are currently documented in unpublished work by Peter Spirtes.

3.2 The nonparanormal distribution

Liu et al. (2009) have extended methods for learning Gaussian graphical models to *nonparanormal* distributions. The nonparanormal includes distributions that can be transformed to multivariate normal distribution over $\mathbf{X}$ by applying a set of monotone functions g to $\mathbf{X}$. Liu et al. (2012, definition 2.1, page 4) define it:

Definition 3.1 (Nonparanormal)**.** Let $g = g_1, \cdots, g_d$ be a set of monotone univariate functions and let $\Sigma \in R^{d \times d}$ be a positive-definite correlation matrix with $\text{diag}(\Sigma) = 1$. A d-dimensional random variable $X = (X_1, \cdots, X_d)^T$ has a nonparanormal distribution $X \sim NPN_d(g, \Sigma)$ if $g(X) := (g_1(X_1), \cdots, g_d(X_d))^T \sim N_d(0, \Sigma)$.

The nonparanormal family is more flexible than the Gaussian, encompassing bimodal and skewed distributions, etc. However, like the Gaussian, the dependence structure among the variables is linear, making it efficient to learn in high dimensions. We can learn graphical models from nonparanormal data without explicitly learning the set of functions g, by applying the nonparanormal SKEPTIC.

The SKEPTIC. Liu et al. (2012) developed the "nonparanormal SKEPTIC".[4] The SKEPTIC estimates Σ, using nonparametric estimators of correlation (see Section 5). PC, FGS, FCI and GFCI all operate directly on $\widehat{\Sigma}$, so we can learn the causal structure without learning $\widehat{g}_v(X)$. Harris and Drton (2013) used the SKEPTIC as the input to PC, and showed that it allowed for accurate causal search on synthetic nonparanormal data. We follow the same approach: we estimate $\widehat{\Sigma}$ using the SKEPTIC, and use that as the input to PC, FGS, FCI and GFCI. We also compare this to using the Pearson correlation matrix.

In reality, transcriptional regulatory networks do not follow a nonparanormal distribution. Interactions between transcription factors are frequently combinatorial rather than additive (Garber et al., 2012). However, the nonparanormal distribution may still be a better approximation than the Gaussian, allowing us to learn some of the regulatory structure while scaling well to high dimensions.

3.3 Genetic regulatory networks

When a gene is expressed, it is first *transcribed* onto mRNA, which may then be *translated* into protein. Gene expression is regulated at the level of transcription, translation, and post-translational modifications of the proteins. Due to the ease of measuring mRNA, we focus on transcriptional regulation. Transcription is regulated by proteins called *Transcription Factors* (TFs) and *chromatin modifiers*. TFs bind to DNA at particular locations and either activate or repress transcription. Chromatin modifications can make DNA more or less accessible to TFs and the transcriptional machinery. We use the following sources of information to validate our models:

ChIP-chip/seq. Chromatin Immuno-Precipitation (ChIP)-chip/seq experiments identify physical interactions between proteins and DNA. If TF A binds to the promoter region of Gene B, we infer that A regulates transcription of B. ChIP-chip/seq experiments are the best method for learning direct regulatory relationships, and have been used for network construction (Boyer et al., 2005; Chen et al., 2008; Hannah et al., 2011; Gerstein et al., 2012; Beck et al., 2013; Xu et al., 2014) as well as for validation. However, ChIP-chip/seq results are extremely sensitive to context and cell type, because transcriptional regulation is likewise sensitive to context and cell type (Neph et al., 2012). We expect both false positives and false negatives: the TF may bind to DNA in vitro which is normally not open for binding in vivo, or vice versa (Ernst and Kellis, 2013); the TF may never detach from a site, so there is no

[4] ↑'Spearman/Kendall Estimates Pre-empt Transformations to Infer Correlation'

regulatory variation in vivo; or the TF binding may be redundant. This makes the interpretation of ChIP-chip/seq studies challenging (DeVilbiss et al., 2014).

LOGOF experiments. 'Loss Or Gain Of Function' (LOGOF) experiments include a variety of manipulations, from knocking out a gene entirely, to transiently increasing its expression. LOGOF experiments are easier to interpret causally than ChIP-chip/seq experiments. However, LOGOF methods do not allow for holding the expression level of other genes constant. They can test whether the manipulated gene is an *ancestor* of other genes, but not whether it is a direct *parent*. This makes validation of the network non-local, a disadvantage compared to ChIP-chip/seq.

3.4 Large-scale evaluations of network construction methods

Constructing Genetic Regulatory Networks (GRNs) from data is an important problem, and many methods have been applied to it. For reviews, see Blais and Dynlacht (2005); Cooke et al. (2009). For comparative evaluations of techniques on synthetic data, see Bansal et al. (2007) and Marbach et al. (2010), and Marbach et al. (2012) for evaluation on real data. Most studies on GRN construction either construct a small network (∼5–30 genes; e.g. Hartemink et al. (2001, 2002)) or test only a few hand-picked causal consequences of the model (e.g. Basso et al. (2005)). To our knowledge, only four studies have attempted to learn a large GRN from expression data and evaluated many causal consequences of the model. We would like to achieve similar or better performance than these four studies. Unfortunately, none of them allow for direct comparison for all algorithms, as described below.

Faith et al. (2007) constructed a regulatory network of *E. coli* using publicly available microarray datasets. They developed the Context Likelihood of Relatedness (CLR) algorithm, which puts an edge between two genes if the corrected pairwise mutual information between those genes is above a threshold. CLR cannot orient edges. Faith et al. (2007) constrained the network so that only TFs could be parents, ruling out many adjacencies and orienting all edges except those between TFs.

Faith et al. (2007) validated their results against RegulonDB, a database of known regulatory relationships in *E. coli* (Salgado et al., 2012). At a precision level of 60%, CLR had ∼6% recall, whereas relevance networks (the next-best algorithm) only had ∼5% recall. However, '60% precision' is likely an overestimate. Precision was only evaluated for edges between genes in RegulonDB. Genes in RegulonDB are more likely to be involved in transcriptional regulation. CLR inferred 1079 edges, but only 338 edges were between pairs of genes in RegulonDB; of those 338, ∼202 corresponded to known effects. The true precision could be as low as 202/1079 = 19% or as high as 943/1079 = 87%. Even using Faith et al. (2007)'s standard for precision, increasing the recall to 10% made precision drop to ∼20%.[5]

[5] ↑Because Faith et al. (2007) don't report the number of edges estimated at other precision levels, we cannot adjust these numbers to take into account the genes not included in RegulonDB.

Marbach et al. (2012) tested many published GRN learning algorithms on data from *E. coli* and *S. cerevisiae*. The gold standard for *E. coli* was RegulonDB, and for *S. cerevisiae* it was a set of ChIP-chip/seq results and regulatory motifs. Success varied dramatically between species. The Area Under the Precision Recall curve (AUPR) was far above chance on *E. coli* data for most algorithms, but none did much better than chance on *S. cerevisiae*. There are at least two plausible explanations: (1) The mechanisms of transcriptional regulation are very different in prokaryotes like *E. coli* than eukaryotes like *S. cerevisiae*; prokaryotic GRNs may simply be easier to learn. (2) RegulonDB may be a better gold standard than the ChIP and motif data used for *S. cerevisiae*.

We cannot directly compare our results to Marbach et al. (2012)'s AUPR scores, because our algorithms output a single graph, corresponding to a point on the PR curve. PC, FGS, FCI and GFCI do include parameters that influence the density of the output network. However, altering these parameters leads PC, FGS, FCI and GFCI to add and remove edges in unpredictable ways. So while we can recover something like a PR curve, the interpretation of the "AUPR" is more difficult.

Cahan et al. (2014) used CLR (Faith et al., 2007) to construct GRNs, which they then used to evaluate cellular reprogramming and directed differentiation protocols. Like Faith et al. (2007) they only allowed edges out of TFs. The GRN construction stage was evaluated on real data from mouse cells, including mESCs, using the ESCAPE database (Xu et al., 2013) as one of the gold standards. We also used this cell type and gold standard. Unfortunately, their results are hard to interpret and are not directly comparable to ours.

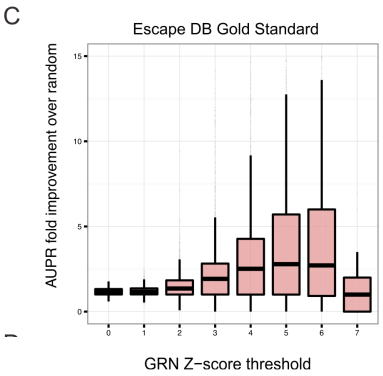


Figure 1: Fig S1 C from Cahan et al. (2014). Original caption: "Combined AUPR based on using the 54 ChIP-ChIP/ChIP-seq transcription factor binding sites in mouse embryonic stem cells from the Escape database as a gold standard."

Cahan et al. (2014) reported CLR's improvement over chance performance at different corrected Z-score thresholds. As a combined measure of precision and recall, they calculated the area of a rectangle $A \times (B - C)$, where A = precision of CLR when $Z = z_i$, B = recall of CLR when $Z = z_i$, and C = recall of CLR when $Z = z_{i+1}$. They called this rectangle the "AUPR", departing from the literature.[6] This area combines an absolute measure of precision with a relative measure of recall: how much the recall for z_i improved over the recall for z_{i+1}. To estimate performance under chance, Cahan et al. (2014) repeatedly sampled random graphs with same number of targets per TF as the CLR graph when $Z = z_i$. For each sampling run r and each z_i, they calculated "AUPR fold improvement over chance", or $\frac{A \times (B-C)}{A_r \times (B_r - C_r)}$. Boxplots of these scores were the final measure of success. Their results on the ESCAPE DB gold standard are reproduced in Fig. 1. It is unclear how to interpret these scores. We present a more natural and principled metric in Section 7.

[6] ↑Personal communication, Patrick Cahan, 4/22/2016.

Maathuis et al. (2010) scored the recovery of *ancestral relationships* in yeast. They estimated the size of the causal effects from the observational data by first learning a GRN structure, then fitting the model. However, PC-Stable returns a Markov equivalence class of models, leading to multiple estimates of the causal effect. Maathuis et al. (2010) used the IDA algorithm (Maathuis et al., 2009) to compute bounds on each causal effect, and ranked all effects according to their lower bounds.

Maathuis et al. (2010) used LOGOF data as the gold standard (Hughes et al., 2000). Ancestral relationships may be weak; Maathuis et al. (2010) considered the largest 10% of effects in the gold standard as the 'true positives'. They reasoned that the chief use of the ranking would be to help biologists prioritize experiments, so they only evaluated performance on the top 5,000 ranked effects (0.4% of effects covered by the gold standard). Results on this subset of effects were very impressive. Precision is greater than 50% for the top 1000 predicted effects.

The IDA algorithm works for CPDAGs but not PAGs, so we perform a similar evaluation for PC-stable and FGS, but not FCI nor GFCI.

4 Approach

Methods. We compare four causal search algorithms: (1) PC-stable, (2) FGS, (3) FCI, and (4) GFCI. We compare two methods of covariance matrix recovery as inputs to these algorithms: (1) Pearson correlation and (2) the nonparanormal SKEPTIC.

Data. We evaluate the methods on single-cell RNA-seq data from mouse embryonic stem cells (mESCs). We incorporate the constraint that only genes known to be transcription factors can be parents. For each algorithm, we run it once with this constraint ("with knowledge") and once without the constraint ("agnostic").

Evaluation. We validate the learned networks against two gold standards: (1) TF-gene interactions from ChIP-x studies, and (2) ancestral relationships from LOGOF studies. We treat the known TF-gene interactions as the true GRN, and evaluate the precision and recall for adjacencies. For orientations, we cannot take the Markov Equivalence Class of the true network because it is cyclic, so we compare the learned orientations against the actual orientations in the gold standard network. We consider the known ancestral relationships to be a set of true effects, and treat their identification as a classification problem. We use IDA (Maathuis et al., 2009) to assign a lower bound on each possible causal effect, then use those bounds to calculate a precision-recall curve for the known ancestral effects.

5 Method

Covariance matrix recovery with the SKEPTIC. We used Pearson correlation and the nonparanormal SKEPTIC to produce two estimates of the covariance matrix. Liu et al. (2012, page 6) describe how the SKEPTIC can use either of two nonparametric estimators of correlation, Spearman's ρ and Kendall's τ. We use the

default implementation of the SKEPTIC in the R package `huge`, which uses Spearman's ρ. Letting r_j^i be the rank of x_j^i among $x_j^1, \cdots, x_j^n$ and $\bar{r}_j = \frac{1}{n}\sum_{i=1}^n r_j^i = \frac{n+1}{2}$, the estimator is:

$$\text{(Spearman's rho)} \qquad \widehat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_j^i - \bar{r}_j)^2 \cdot \sum_{i=1}^n (r_k^i - \bar{r}_k)^2}}$$

Liu et al. (2012, page 7) connect Spearman's rho and Kendall's tau to Σ using a lemma from (Kendall, 1948) and (Kruskal, 1958):

Lemma 5.1. *Assuming* $X \sim NPN_d(f, \Sigma)$*, we have* $\Sigma_{jk} = 2\sin(\frac{\pi}{6}\rho_{jk}) = \sin(\frac{\pi}{2}\tau_{jk})$.

Liu et al. (2012) then use ρ to estimate Σ: $\widehat{\Sigma}^{\rho}_{jk} = 2\sin(\frac{\pi}{6}\widehat{\rho}_{jk})$ when $j \neq k$ and $\widehat{\Sigma}^{\rho}_{jk} = 1$ when $j = k$.

Causal network recovery. For all causal search algorithms, we use the most recent implementation in the Tetrad software package (Ramsey et al., 2015). Tetrad source code and compiled jars are available from the project's Git repository. We used the algorithms PC-Stable, FGS, FCI and GFCI. For PC-Stable, FCI and GFCI, we set the maximum conditioning set size (the 'depth') to 3, and for FCI and GFCI we constrained the length of the longest inducing path used to orient edges (the 'maxPathLength') to 3 to reduce runtime. We tried several values of the penalty discount for FGS and GFCI, and the α value for PC and FCI, and chose penaltyDiscount $= 15$ and $\alpha = 10^{-5}$ as these produced sparse graphs with decent precision.

To make our results comparable to Faith et al. (2007) and Cahan et al. (2014)'s, we compared search without background knowledge ('agnostic') to search constrained such that edges were only allowed out of TFs.

6 Data

6.1 Data for network learning

We used a dataset of gene expression levels from 934 mouse embryonic stem cells. Klein et al. (2015) originally collected this dataset to demonstrate their microfluidics-based single-cell RNA-seq procedure, 'inDrop'. Most single-cell RNA-seq protocols are extremely noisy, because they measure tiny samples of RNA, which undergo successive rounds of exponential amplification. InDrop reduces noise two ways: (1) linear amplification in the first round, and (2) by attaching a unique molecular identifier (UMI) to each mRNA transcript in the first round. Klein et al. (2015) filtered out amplified noise by counting the *distinct UMIs* for each transcript rather than the total number of transcripts. The large sample size and low noise make Klein et al. (2015)'s dataset ideal for network reconstruction.

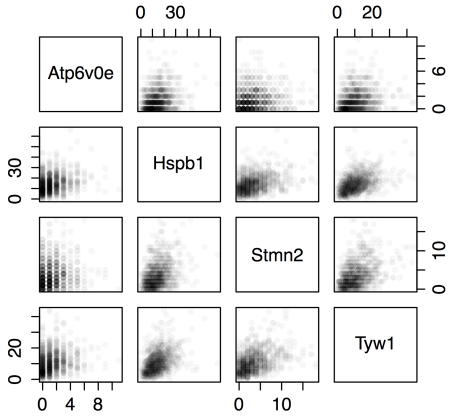


Figure 2: Bivariate plots of genes Atp6v0e, Hspb1, Stmn2, and Tyw1.

Of the 24,175 measured genes, Klein et al. (2015) identified 2047 genes with significant variance, using a false discovery rate cutoff of 0.10. To reduce runtime, we limited network reconstruction to the 465 highest variance genes, plus another 35 genes within the top 2047 that code for transcription factors. This gave us a final variable set of 500 genes, 120 of which were TFs or chromatin modifiers.

Figure 2 shows bivariate plots of four randomly selected genes. It shows that gene expression is (a) heteroskedastic, and (b) integer valued and non-negative, corresponding to counts of transcripts. Heteroskedasticity indicates that the nonparanormal is more appropriate than the Gaussian. Integer values, however, imply that the data are not nonparanormal. The nonparanormal is only an approximation.

One might worry that the non-negativity would bias estimates of correlation (Regier and Hamdan, 1971). We checked this, using Nie et al. (2008)'s method to correct Kendall's τ for truncation bias. We ran each graph-learning algorithm on three correlation matrices: Pearson correlation, the SKEPTIC (using Spearman's ρ), and the truncation-corrected SKEPTIC (using Kendall's τ). Truncation-correction did not improve performance in any condition.

6.2 Data for evaluation

Our knowledge of the true network is incomplete, so apparent 'false positives' may in fact be novel results. Furthermore, regulatory relationships are highly context-dependent. The ideal gold standard data would be collected in exactly the same cell type under the same conditions as our observational data. Because we did not collect our own data, we settled for close matches.

ChIP-chip/seq data for direct edge evaluation came from two sources:

- Xu et al. (2013)'s ESCAPE database contains 107,980 non-redundant TF-gene interactions learned from ChIP-chip/seq experiments in mouse embryonic stem cells. 1,921 of these interactions were between genes in our 500 genes of interest.
- TF-gene interactions compiled by Neph et al. (2012) and Stergachis et al. (2014) were downloaded from `http://www.regulatorynetworks.org/`. [7] Combined, these lists included 42,486 non-redundant interactions, of which 177 were between genes in our set of 500.

The union of the two gold standards included 2,098 known TF-gene interactions.

Mutation & expression manipulation for path evaluation. Our gold standard data on ancestral relationships likewise came from multiple sources:

- Xu et al. (2013)'s ESCAPE database contains 101,673 non-redundant regulatory relationships learned from LOGOF experiments in mESCs, of which 1,778 interactions were between genes in our top 500.

[7] ↑The lists of interactions were drawn from three files with mouse ZhBTc4 embryonic stem cells (control, +6 hours doxycycline, +24 hours doxycycline), and one with mouse mCj7 embryonic stem cells (via 129S1/SVImJ mice).

- Correa-Cerro et al. (2011) engineered a set of mESC cell lines for inducible expression of particular TFs. They learned 140,157 non-redundant regulatory relationships, of which 892 were between genes in our top 500.
- Nishiyama et al. (2013) used shRNAs to systematically repress the expression of several TFs in mESCs. They learned 10,223 relationships causing a significant change in expression. Of these, 275 were between genes in our top 500.

Due to overlap, the union of the LOGOF gold standards included 1,767 regulatory relationships. (Self-regulatory relationships were excluded.)

6.3 Background knowledge

Any genes listed as being parents or ancestors of other genes in any of our gold standards were considered TFs. We used the Gene Ontology (Carbon et al., 2009) to identify additional TFs and chromatin modifiers. Genes annotated with 'GO:0016568 : chromatin modification' or 'GO:0006355 : regulation of transcription, DNA-templated' were considered TFs. Of our 500 genes of interest, 120 genes were in the list of TFs.

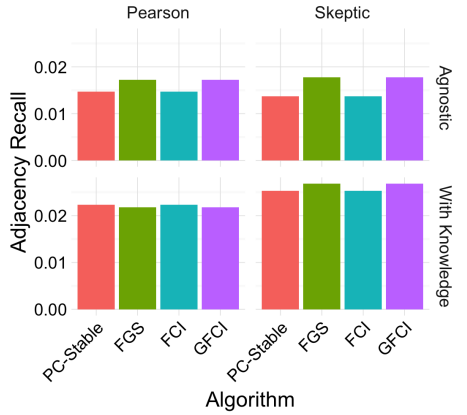
7 Results

Figure 3 shows results on the ChIP-chip/seq gold standard. Using heavy penalties on search ($\alpha = 10^{-5}$ for PC-stable and FCI, and penalty = 15 for FGS and GFCI) we produced sparse graphs (Fig. 3c) with very low recall (Fig. 3a) and only moderate precision (Fig. 3b) for adjacencies. However, all algorithms performed highly significantly better than chance, according to a hypergeometric test (Fig. 3d).[8] Using the SKEPTIC with background knowledge, FGS and GFCI identified 53 true adjacencies out of 235 estimated edges, a 14-fold improvement over the number of successes expected under chance ($p < 10^{-45}$). This 22.6% precision may be high enough to help prioritize experiments, in combination with other sources of knowledge.

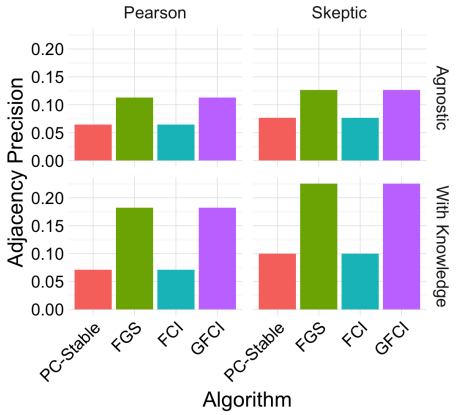
FGS and GFCI have slightly better recall of adjacencies and much better precision than PC-stable and FCI, indicating that score-based algorithms perform better than constraint-based algorithms on these data. FGS and GFCI perform equally well, as do PC-stable and FCI; relaxing the assumption of no latent variables does not improve performance. Background knowledge improves performance in every condition. Using the SKEPTIC covariance matrix instead of the Pearson matrix improves precision and recall of adjacencies when background knowledge is available, but has only a slight effect without background knowledge.

Because the ChIP-chip/seq gold standard graph is cyclic, we cannot use its Markov equivalence class to evaluate orientations. Instead we compare estimated

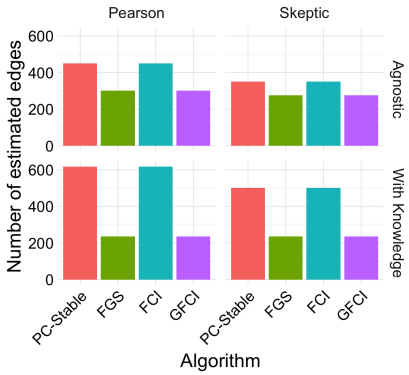
[8] ↑By 'chance' we mean choosing a random graph of equal density to the estimated graph. That is, drawing k adjacencies without replacement, where k is the number of edges in the estimated graph, out of a set of N edges (the number of possible adjacencies) that includes n successes (the number of true adjacencies in the gold standard). The hypergeometric is the null distribution.



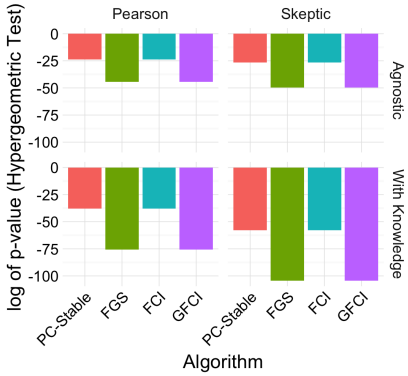
(a) Recall: $\frac{\#\text{correct adjacencies}}{\#\text{true adjacencies in gold standard}}$



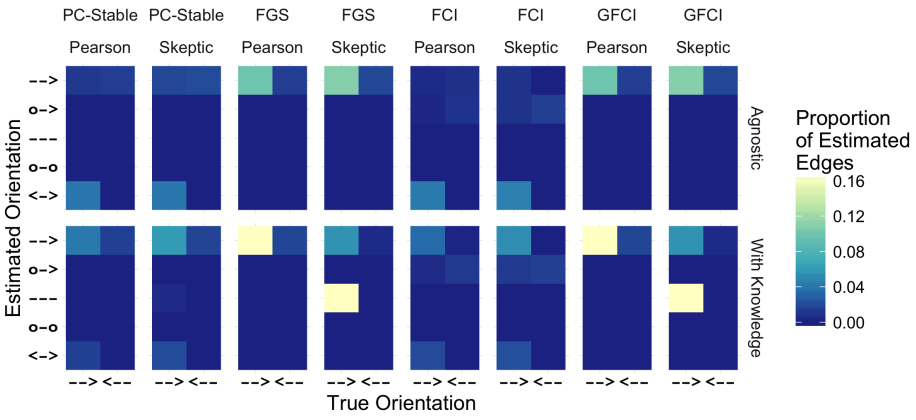
(b) Precision: $\frac{\#\text{correct adjacencies}}{\#\text{estimated adjacencies}}$



(c) # Estimated adjacencies.



(d) Log p-value of hypergeometric test.



(e) Normalized orientation confusion matrices: $\frac{\#\text{edges w/ that est'd \& true orientation}}{\#\text{estimated edges}}$ (the denominator includes false positives, i.e. cases where 'True Orientation = No edge', although this column is not shown). Ideally all edges would be in the top left, where estimated orientation matches the truth.

Figure 3: Results on the ChIP-chip/seq gold standard: precision and recall for adjacencies, confusion matrix for orientations.

orientations to the actual directed edges from the gold standard. The orientation confusion matrix (Fig. 3e) also shows FGS and GFCI outperforming PC-stable and FCI. However, when background knowledge is available, the SKEPTIC *hurts* FGS and GFCI's performance: it produces lots of undirected edges, whereas the Pearson matrix produces mostly directed edges and has the best precision for orientations.

Results on the LOGOF gold standard were less impressive.[9] Maathuis et al. (2010) looked at the top 0.4% of their ROC curve; our gold standard is much smaller, so we look at the top 10%. FGS with background knowledge and the SKEPTIC achieved the best result, a partial AUC of 51.7% (Fig. 4). By comparison, PC-Stable using the Pearson matrix and background knowledge achieved a pAUC of 51.1% (Fig. 5; PC-Stable performed worse with the SKEPTIC).

Using PC-stable on yeast microarray data, Colombo and Maathuis (2012) achieved ∼50% precision in the top 1000 effects. Using the same algorithm plus background knowledge, our precision in the top 1000 effects was 12%. It seems that this dataset and gold standard are not as favorable as those used by Maathuis et al. (2010) and Colombo and Maathuis (2012).

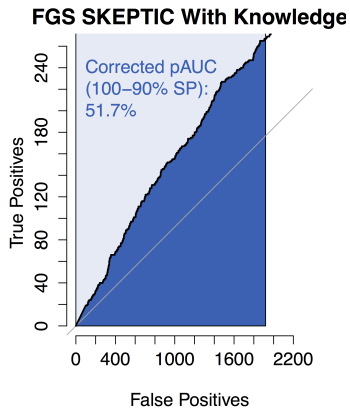


Figure 4: Partial AUC in top 10% of ROC curve for IDA using FGS, the SKEPTIC matrix, and background knowledge, compared to the LOGOF gold standard.

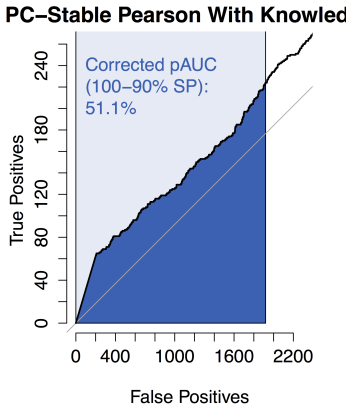


Figure 5: Partial AUC in top 10% of ROC curve for IDA using PC-stable, the Pearson matrix, and background knowledge, compared to the LOGOF gold standard.

8 Discussion

Our absolute performance cannot be compared to any previous study. Performance varies dramatically either by species, or by gold standard, as demonstrated by Marbach et al. (2012) and again by the comparison of Colombo and Maathuis (2012)'s results with our results using IDA and PC-Stable. Only Cahan et al. (2014) used the same species and gold standard as us, and they used a non-standard evaluation metric.

Most previous results are reported as AUPR values, which our algorithms cannot generate. We can compare our results to a point on the precision-recall curve published by Faith et al. (2007). Our best result was 2.7% recall and 22.6% precision, which is strictly dominated by Faith et al. (2007)'s 6% recall and 60% precision. However, it is unclear whether this difference is due to our algorithms' inferiority. Marbach et al. (2012) showed that CLR performed much worse on *S. cerevisiae* than *E. coli*, so we may have simply used a more challenging species or a less reliable gold standard. By limiting evaluation to pairs of genes in RegulonDB, Faith et al. (2007) may have overestimated precision, as described in Section 3.4.

[9] ↑Note that FCI and GFCI cannot be evaluated by this standard because IDA has not yet been extended to work with PAGs, so results are only evaluated for FGS and PC-Stable.

We confirmed our hypothesis that the SKEPTIC could improve performance: in the presence of background knowledge it improved precision for adjacencies, although it hurt for orientations. However, to our surprise, allowing the presence of latent variables (by using FCI or GFCI) did not improve performance.

This study sets the first benchmark for learning GRNs from single-cell RNA-seq data. The hypergeometric test is a natural and interpretable test of success. It shows that all our algorithms perform unambiguously better than chance at identifying adjacencies (see Fig. 3d; for FGS with the SKEPTIC and background knowledge, $p < 10^{-45}$). This is the first time a large-scale network construction algorithm has been shown to be unabiguously better than chance at detecting adjacencies in a eukaryotic GRN, without incorporating additional data from intervention distributions.

The results for learning ancestral relationships, evaluated by the intervention gold standard, are less impressive than those demonstrated by Maathuis et al. (2010) and Colombo and Maathuis (2012). Even using the same algorithm as Colombo and Maathuis (2012), we achieved little elevation at the start of the ROC curve. This may be because our interventional gold standard was too small or low quality.

9 Limitations

Learning eukaryotic GRNs from observational data is an extremely difficult problem. This study has promising results but suffers from a number of limitations.

Our results are not comparable to the results of previous studies, for two reasons: (1) In order to explore the potential of single-cell RNA-seq data, we used a new dataset, necessitating the use of a different gold standard; we did not reimplement the methods from similar studies, so any differences may be due to the new methods or the new data. (2) Marbach et al. (2012) performed a comprehensive evaluation of methods, but reported only the AUPR for each method; our methods do not produce a precision-recall curve for adjacencies (although IDA can produce such a curve for ancestral relationships).

Our observational data and gold standard data were collected under different conditions; the gold standards were collected by multiple labs. Gene regulation is highly context sensitive, so these differences matter. Furthermore, ChIP-chip/seq results are not perfect indicators of causal relationships; even in ideal conditions this "gold standard" includes both false positives and false negatives.

Single-cell RNA-seq has low sampling efficiency, leading to low estimates of transcript abundance. This is particularly problematic for TFs, which typically exist in low concentrations.

The nonparanormal distribution is an imperfect approximation to the distribution of mRNA counts. It can only model linear dependencies between variables. There is evidence that transcriptional regulation is non-linear, with multiple transcription factors interacting to turn expression on or off.

10 Conclusions

Using the SKEPTIC with FGS or GFCI produced the best results for learning adjacencies, whereas the Pearson correlation produced the best results for orientations. Researchers may find a combined approach most useful. Score-based algorithms outperformed constraint-based algorithms. Allowing the presence of latent variables did not improve performance, but restricting search so that edges were only allowed out of TFs improved performance for all algorithms.

Causal search algorithms identified adjacencies 14× times better than chance when using real single-cell RNA-seq data from mouse embryonic stem cells. This is a promising result, and sets a benchmark for future studies attempting network reconstruction from single-cell RNA-seq data. We identified adjacencies with a precision of 22.6%, which may help researchers prioritize experiments.

11 Future research

There are several ways future research could build upon this result. Modeling gene expression as a Poisson process would capture the fact that the measurements are counts. Incorporating intervention data, using the techniques developed by Mordelet and Vert (2008), Qin et al. (2014) and Shojaie et al. (2013), could improve performance substantially. Collaborating with wet lab to produce gold standard data and observational data collected in the same cell type, under the same conditions, would dramatically reduce the noise in evaluation. Reimplementing previously published algorithms and applying them to this dataset, and evaluating with this gold standard, would help determine whether our promising results are due to using single-cell data, or different algorithms. Extending our current methods so that they produce a PR curve would allow us to compare with the AUPRs in previous publications. We hope to see future studies taking up these challenges.

12 Acknowledgments

I wish to thank Peter Spirtes, David Danks, Joe Ramsey, Roy Maxion and Dana Pe'er for their invaluable help and guidance.

This research was supported by the Department of Philosophy at Carnegie Mellon University, by grant DARPA-BAA-14-14 awarded by DARPA, and by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The content is solely the responsibility of the author and does not necessarily represent the official views of DARPA, the NIH or the CMU Philosophy Department.

References

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1). 6

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature genetics*, 37(4):382–390. 6

Beck, D., Thoms, J. A., Perera, D., Schütte, J., Unnikrishnan, A., Knezevic, K., Kinston, S. J., Wilson, N. K., O'Brien, T. A., Göttgens, B., et al. (2013). Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood*, 122(14):e12–e22. 5

Blais, A. and Dynlacht, B. D. (2005). Constructing transcriptional regulatory networks. *Genes & development*, 19(13):1499–1511. 6

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956. 5

Cahan, P., Li, H., Morris, S. A., da Rocha, E. L., Daley, G. Q., and Collins, J. J. (2014). Cellnet: network biology applied to stem cell engineering. *Cell*, 158(4):903–915. 7, 9, 13

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Group, W. P. W., et al. (2009). Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289. 11

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117. 5

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554. 4

Chickering, D. M. and Meek, C. (2002). Finding optimal bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligenceconference on Uncertainty in artificial intelligence*, pages 94–102. Morgan Kaufmann Publishers Inc. 4

Chu, T., Glymour, C., Scheines, R., and Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152. 1

Colombo, D. and Maathuis, M. H. (2012). Order-independent constraint-based causal structure learning. *arXiv preprint arXiv:1211.3295*. 2, 3, 13, 14

Cooke, E. J., Savage, R. S., and Wild, D. L. (2009). Computational approaches to the integration of gene expression, ChIP-chip and sequence data in the inference of gene regulatory networks. *Seminars in cell & developmental biology*, 20(7):863–868. 6

Correa-Cerro, L. S., Piao, Y., Sharov, A. A., Nishiyama, A., Cadet, J. S., Yu, H., Sharova, L. V., Xin, L., Hoang, H. G., Thomas, M., et al. (2011). Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Scientific reports*, 1(167):DOI:10.1038/ srep00167. 11

DeVilbiss, A. W., Sanalkumar, R., Johnson, K. D., Keles, S., and Bresnick, E. H. (2014). Hematopoietic transcriptional mechanisms: from locus-specific to genome-wide vantage points. *Experimental hematology*, 42(8):618–629. 6

Ernst, J. and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome research*, 23(7):1142–1154. 5

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLOS Biology*, 5(1):0054–0066. 6, 7, 9, 13

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular cell*, 47(5):810–822. 5

Geiger, D. and Pearl, J. (1988). On the logic of causal models. In *Proceedings of the Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*, pages 136–147, Corvallis, Oregon. AUAI Press. 21

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100. 5

Hannah, R., Joshi, A., Wilson, N. K., Kinston, S., and Göttgens, B. (2011). A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Experimental hematology*, 39(5):531–541. 5

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research*, 14(1):3365–3383. 2, 5

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, pages 437–449. 6

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., Young, R. A., et al. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pacific symposium on biocomputing*, volume 6, pages 422–433. 6

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126. 8

Kendall, M. G. (1948). *Rank correlation methods.* Griffin, Oxford, England. 9

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201. 1, 9, 10

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861. 9

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326. 2, 4, 5, 8, 9

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328. 4

Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248. 8, 13, 14

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164. 2, 8

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Stolovitzky, G., et al. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804. 6, 7, 13, 14

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291. 6

Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–418, San Francisco, CA. Morgan Kaufmann. 21

Mordelet, F. and Vert, J.-P. (2008). SIRENE: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82. 15

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286. 5, 10

Nie, L., Chu, H., and Korostyshevskiy, V. R. (2008). Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits. *Canadian Journal of Statistics*, 36(3):427–442. 10

Nishiyama, A., Sharov, A. A., Piao, Y., Amano, M., Amano, T., Hoang, H. G., Binder, B. Y., Tapnio, R., Bassey, U., Malinou, J. N., et al. (2013). Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Scientific reports*, 3. 11

Pearl, J. (2009). *Causality*. Cambridge university press, 2nd edition. 1

Pearl, J., Verma, T., et al. (1991). *A theory of inferred causation*. Morgan Kaufmann San Mateo, CA. 1

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053. 21

Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., and Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 67(3):294–303. 15

Ramsey, J. D. (2015). Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*. 4

Ramsey, J. D., Spirtes, P., Glymour, C., and Scheines, R. (2015). Tetrad v. `http://www.phil.cmu.edu/tetrad`. 9

Regier, M. H. and Hamdan, M. (1971). Correlation in a bivariate normal distribution with truncation in both variables. *Australian Journal of Statistics*, 13(2):77–82. 10

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461. Morgan Kaufmann Publishers Inc. 3

Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J. S., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, A., Porrón-Sotelo, L., Huerta, A. M., Bonavides-Martínez, C., Balderas-Martínez, Y. I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., Moral-Chávez, V. D., Hernández-Alvarez, A., Morett, E., and Collado-Vides, J. (2012). Regulondb (version 8.0): Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, doi: 10.1093/nar/gks1201

PMID: 23203884 PMC: PMC3531196. 6

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030. 21

Shojaie, A., Jauhiainen, A., Kallitsis, M., and Michailidis, G. (2013). Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles. *arXiv preprint arXiv:1312.0335*. 15

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. The MIT Press, Cambridge, MA., 2nd edition. 1, 3, 4, 21

Spirtes, P., Glymour, C. N., and Scheines, R. (1989). Causality from probability. Technical Report CMU-LCL-89-4, Department of Philosophy, Carnegie Mellon University. 1

Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., et al. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–370. 10

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752. 1

Xu, H., Ang, Y.-S., Sevilla, A., Lemischka, I. R., and Ma'ayan, A. (2014). Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Computational Biology*, 10(8):e1003777. 5

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E. Y., Tan, C. M., Kou, Y., Kim, Y. E., Lemischka, I. R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database*, 2013:bat045. 7, 10

Zhang, J. (2008). Causal reasoning with ancestral graphs. *The Journal of Machine Learning Research*, 9:1437–1474. 22

Zhang, J. and Spirtes, P. (2003). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the nineteenth conference on uncertainty in artificial intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc. 21

Zhang, J. and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271. 21

13 Appendix: Background on Causation

13.1 Causal Markov Assumption

The Causal Markov Assumption (CMA) links the graphical structure to the independence structure of the probability distribution over X. By the CMA,

Definition 13.1 (Causal Markov Assumption)**.** A variable X_i is independent of its non-descendants given its direct parents.

The CMA allows us to read off the independence relationships between variables from the structure of the graph. However, the CMA is only a reasonable assumption when Causal Sufficiency holds:

Definition 13.2 (Causal Sufficiency)**.** A set of variables V is 'causally sufficient' iff for every variable k such that k is a parent of two distinct variables $v, v' \in V$, then $k \in V$.

For causal sufficiency to hold, no confounding variables can be omitted from the causal graph. One important consequence of the CMA is the factorization theorem:

Theorem 13.1 (Factorization Theorem)**.** *Given a DAG G and a corresponding probability distribution P_G, if P_G satisfies the Causal Markov Assumption relative to G, then P_G factors into $\prod_v (X_v | Pa(X_v))$, where $Pa(X_v)$ are the parents of v in G.*

However, we may want to know about independences conditional on some other set of variables. In *acyclic* directed graphs, a complete characterization of conditional independences implied by the model is given by d-separation.

13.2 d-separation

The full set of conditional independences implied by a DAG are characterized by d-separation. If P_G satisfies the CMA relative to G, then $X_i \perp\!\!\!\perp X_j | S$ if X_i is d-separated from X_j conditional on S.

Definition 13.3 (d-separation)**.** X_i is d-separated from X_j conditional on S iff there is no active path between X_i and X_j in G conditional on S. A path is *active* iff every node on it is active. A node v is active on a path p if either:

1. v is a non-collider on p, and $v \notin S$, or
2. v is a collider on p, and either $v \in S$, or there is some descendant $q \in Descendants(v)$ such that $q \in S$

where a *collider* on a path p is a non-endpoint node k such that the two edges in p that include k are both oriented into k.

13.3 Faithfulness

The Causal Markov Condition implies that if two variables are dependent in the probability distribution, they must be d-connected in the causal graph. However, it does not tell us which variables should be d-separated in the graph. We could always satisfy Markov by returning a complete graph. In order to learn sparse models from data, we must make an additional assumption about the relationship between the distribution and the graph.

Definition 13.4 (Faithfulness)**.** A distribution P is *faithful* to a graph G if: for every pair of variables $X_i \neq X_j$ and set $\mathbf{S}$, if X_i and X_j are d-connected given $\mathbf{S}$ in G, then X_i is dependent on X_j conditional on $\mathbf{S}$ in P.

Faithfulness means that every d-connection in the graph gives rise to a probabilistic dependence; so every independence in the distribution must correspond to a d-separation in the graph. When we see a conditional independence in the data, we can infer a separation in the graph. Faithfulness is a strong assumption. Some causal search algorithms rely on weaker assumptions of minimality or triangle-faithfulness – see Spirtes et al. (2000); Zhang and Spirtes (2008) for more details. Faithfulness can also be strengthened to provide guarantees of uniform consistency for causal learning (Zhang and Spirtes, 2003).

13.4 DAGs vs. CPDAGs vs. MAGs vs. PAGs

We use different kinds of graphs to represent the causal system vs. what we can *learn* about the causal system. Table 2 summarizes the relationships between these models.

	Individual model	Markov equivalence class of models
Causally sufficient	DAG	CPDAG
Causally insufficient	MAG	PAG

Table 2: Relationships between DAGs, CPDAGs, MAGs and PAGs

Observed dependences and independences sometimes allow us to distinguish between causal models. For example, if the graph G has $V = \{A, B\}$, and we observe that $A \not\perp\!\!\!\perp B$, this rules out the model in which A is d-separated from B, by the Causal Markov Assumption. However, we cannot distinguish between the models $A \to B$ and $A \leftarrow B$.

The set of models that entail the same conditional independences as G are called the *Markov Equivalence Class* (MEC) of G. For multivariate Gaussian and multinomial data, one can only distinguish between causal models using conditional independences (Geiger and Pearl, 1988; Meek, 1995).[10] As a result, when we evaluate causal search algorithms, we evaluate how well they recover the MEC of the true model.

The MEC of a DAG can be represented by a Completed Partially Directed Acyclic Graph or CPDAG. All members of the MEC have the same adjacencies and the same V-structures. A CPDAG therefore has the same adjacencies as the member graphs it represents. If an edge is oriented the same way in every member of the MEC, it is directed in the CPDAG; otherwise it is undirected.

We may wish to relax the assumption of causal sufficiency. We can make a weaker assumption that there *exists* some causally sufficient graph with respect to which our

[10] ↑Most other distributions include additional information (Shimizu et al., 2006; Peters et al., 2014), allowing us to further distinguish within the Markov equivalence class. However, algorithms leveraging this information are computationally intense, do not scale well to high dimensions, and do not perform well if we relax the assumption of causal sufficiency.

distribution is Markov. In that case, we can still represent the *ancestral* relationships between the observed variables, using a Maximal Ancestral Graph (MAG). A MAG allows that there may be additional unobserved confounders. Each edge in a MAG corresponds to two causal statements, one for each endpoint of the edge. A tail endpoint denotes "is an ancestor of" wherease an arrow denotes "is not an ancestor of", so the edge $A \leftrightarrow B$ represents the statement that neither A nor B is the ancestor of the other; and because they are connected, we can infer that they must both be children of an unobserved confounder.

Just like DAGs, there are Markov equivalence classes of MAGs: sets that cannot be distinguished by conditional independences alone. We can represent the MEC of a MAG using a Partial Ancestral Graph or PAG. The PAG represents everything we can learn about the causal structure from conditional independences. For more on the interpretation of MAGs and PAGs, see Zhang (2008).