The Annals of Applied Statistics
2019, Vol. 13, No. 2, 848–873
https://doi.org/10.1214/18-AOAS1213

GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION

BY ANDREW MCDAVID[*,1], RAPHAEL GOTTARDO[†,‡,1,2], NOAH SIMON[‡] AND MATHIAS DRTON[‡,§,3]

University of Rochester Medical Center[*], Fred Hutchinson Cancer Research Center[†], University of Washington[‡] and University of Copenhagen[§]

Bulk gene expression experiments relied on aggregations of thousands of cells to measure the average expression in an organism. Advances in microfluidic and droplet sequencing now permit expression profiling in single cells. This study of cell-to-cell variation reveals that individual cells lack detectable expression of transcripts that appear abundant on a population level, giving rise to zero-inflated expression patterns. To infer gene coregulatory networks from such data, we propose a multivariate Hurdle model. It is comprised of a mixture of singular Gaussian distributions. We employ neighborhood selection with the pseudo-likelihood and a group lasso penalty to select and fit undirected graphical models that capture conditional independences between genes. The proposed method is more sensitive than existing approaches in simulations, even under departures from our Hurdle model. The method is applied to data for T follicular helper cells, and a high-dimensional profile of mouse dendritic cells. It infers network structure not revealed by other methods, or in bulk data sets. A R implementation is available at https://github.com/amcdavid/HurdleNormal.

1. Introduction. Graphical models have been used to synthesize high-throughput gene expression experiments into understandable, canonical forms (Dobra et al. (2004), Markowetz and Spang (2007)). Although inferring causal relationships between genes is perhaps the ultimate goal of such analysis, causal models may be difficult to estimate with observational data, and experimental manipulation of specific genes has remained costly, and largely inimitable to high-throughput biology. Many analyses have thus focused on undirected graphical models (also known as Markov random fields) that capture the conditional independences present between gene expression levels. The graph determining such a model describes each gene's statistical predictors: each gene is optimally predicted using only its neighbors in the graph. With gene expression studies serving as key

Received October 2016; revised March 2018.

[1]Supported by grant R01 EB008400 from the National Institute of Biomedical Imaging and Bioengineering, US National Institutes of Health.

[2]Supported by a grant from the Bill and Melinda Gates foundation, the Vaccine and Immunology Statistical Center (VISC), OPP1032317.

[3]Supported in part by NSF Grant DMS-1561814.

Key words and phrases. Gene network, single cell gene expression, graphical model, group lasso.

motivation, a host of different approaches have been developed for structure learning and parameter estimation in undirected graphical models (Drton and Maathuis (2017)).

Characterization of the conditional independences between genes answers a variety of scientific questions. It can help falsify models of gene regulation, since statistical dependence is expected, given causal dependence. In immunology, polyfunctional immune cells, which simultaneously and nonindependently express multiple cytokines, are useful predictors of vaccine response (Precopio et al. (2007)). Simultaneous expression or *co-expression* of cellular surface markers potentially define new cellular phenotypes (Lin et al. (2015)), so expanding the "dictionary" of co-expression allows phenotypic refinements. Graphical models allow one to study such co-expression at the level of direct interactions.

1.1. *Single cell gene expression*. Established technology determines gene expression levels by assaying bulk aggregates of cells assayed through microarrays or RNA sequencing. Although graphical modeling of the resulting data has seen profitable applications (see, e.g., Li, Pearl and Jackson (2015)), there is an inheritant limitation to what can be inferred from expression levels that are averages across hundreds or thousands of individual cells, as we discuss in Section 2. In contrast, recent microfluidic and molecular barcoding advances have enabled the measurement of the minute quantities of mRNA present in *single cells*. This new technology provides a unique resolution of gene co-expression and has the potential to facilitate more interpretable conclusions from multivariate data analysis and, in particular, graphical modeling.

At the same time, single cell expression experiments bring about new statistical challenges. Indeed, a distinctive feature of single cell gene expression data—across methods and platforms—is the bimodality of expression values (Finak et al. (2015), Marinov et al. (2014), Shalek et al. (2014)). Genes can be "on," in which case a positive expression measure is recorded, or they can be "off," in which case the recorded expression is zero or negligible. Although the cause of this *zero-inflation* remains unresolved, its properties are of intrinsic interest (Kim and Marioni (2013)). It has been argued that the zero-inflation represents censoring of expression below a substantial limit of detection, yet comparison of *in silico* signal summation from many single cells, to the signal measured in biological sums of cells suggest that the limit of detection is negligible (McDavid et al. (2013)). Moreover, the empirical distribution of the log-transformed counts appears rather different than would be expected from censoring: the distribution of the log-transformed, positive values is generally symmetric. Yet the presence of bimodality in technically replicated experiments ("Pool/split" experiments) implicates the involvement of technical factors (Marinov et al. (2014)).

Zero-inflation is seen, in particular, in a single cell gene expression experiment we analyze in Section 6. The experiment concerns T follicular helper (Tfh) cells, which are a class of $CD4^+$ lymphocytes. B-cells that secrete antibodies require Tfh cell co-stimulation to become active (Ma et al. (2012)). Tfh cells are defined, and

identified both through their location in the B-cell germinal centers, as well as their production of high levels of the proteins CXCR5, PD1 and BcL-6. In the experiment we consider, Tfh cells were identified from cells from lymph node biopsy producing protein for $CD4^{+}CXCR5^{+}PD1^{+}$. Figure 1 shows the pairwise RNA expression distribution of four Tfh marker genes ($P < 10^{-20}$ compared to non-Tfh lymph node T-cells, which are not shown). Although the expression of these genes

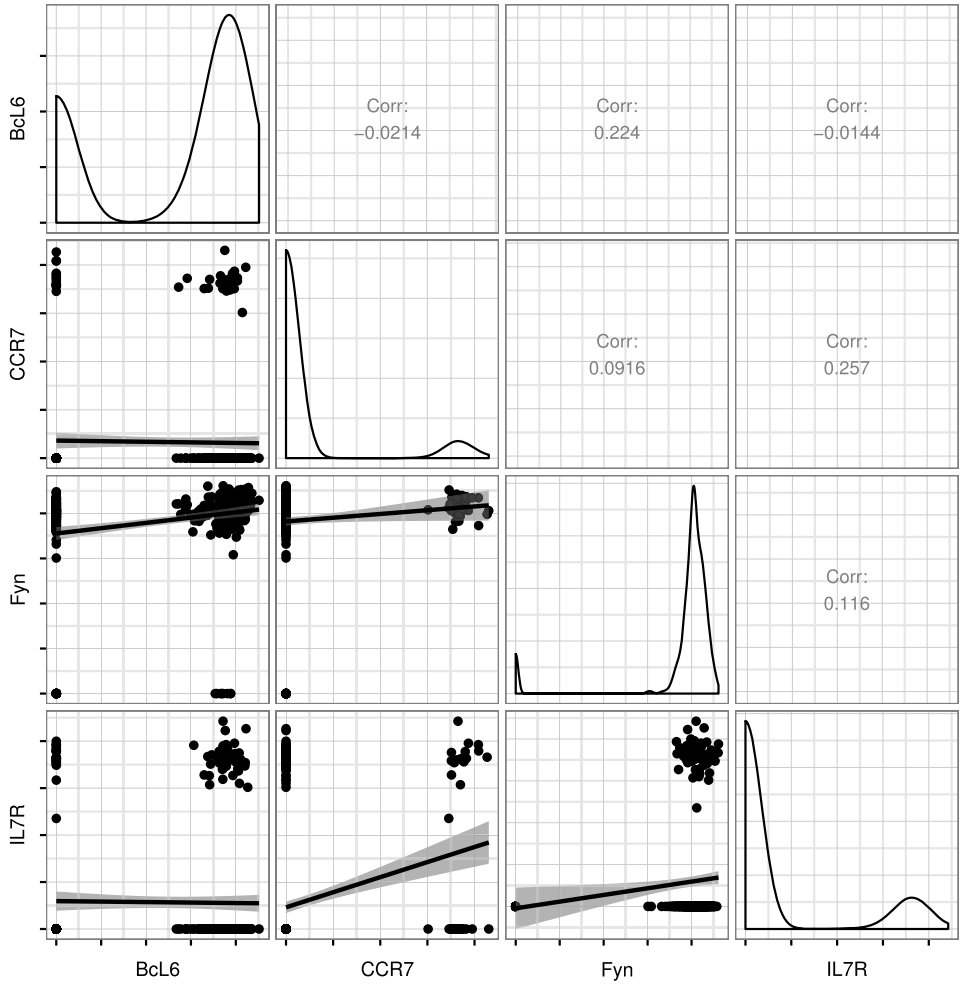

FIG. 1. *Scatter plots of inverse cycle threshold* (40-Ct) *measurements from a quantitative PCR* (*qPCR*)*-based single cell gene expression experiment* (*lower panels*). *The cycle threshold* (*Ct*) *is the PCR cycle at which a predefined fluorescence threshold is crossed, so a larger inverse cycle threshold corresponds to greater log-expression* (*McDavid et al.* (2013)). *Measurements that failed to cross the threshold after* 40 *cycles are coded as* 0. *Marginal expression in Tfh* ($CXCR5^{+}PD1^{+}$) *cells of Tfh marker genes is illustrated in the kernel-density estimates along the diagonal. The lower panels show the linear relationships between pairs of genes.*

could help discriminant Tfh from non-Tfh cells, the strength of linear relationships within Tfh cells (upper panels) varies. To identify co-expressing subsets of cells or to clarify the conditional relationship between genes, estimating the multivariate dependence structure of expression within Tfh cells is necessary. Figure 1 illustrates the issue of zero-inflation. The data are clearly poorly modeled by the linear regression models whose fit is shown in the lower panels of the figure.

1.2. *Modeling zero-inflation*. In order to accommodate the distributional features observed in single cell gene expression, we propose a joint probability density function $f(\mathbf{y})$ of the form

$$\log f(\mathbf{y}) = \mathbf{v_y}^T \mathbf{G} \mathbf{v_y} + \mathbf{v_y}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C(\mathbf{G}, \mathbf{H}, \mathbf{K}), \qquad \mathbf{y} \in \mathbb{R}^m, \tag{1}$$

for the dominating measure obtained by adding a Dirac mass at zero to the Lebesgue measure. The vector $\mathbf{y} \in \mathbb{R}^m$ comprises the expression levels of m genes in a single cell, and the vector $\mathbf{v_y} \in \{0, 1\}^m$ is defined through element-wise indicators of nonzero expression, so $[\mathbf{v_y}]_i = I_{\{y_i \neq 0\}}$ for $i = 1, \ldots, m$. In the specification from (1), both binary and continuous versions of gene expression are sufficient statistics, and interactions thereof are parametrized, with $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{K}$ being matrices of interaction parameters. Zeros in these interaction matrices indicate conditional independences (and thus, absence of edges in a graph for a graphical model). Specifically, the ith and jth coordinate are conditionally independent if and only if all interaction matrices have their (i, j) and (j, i) entries zero (Lauritzen (1996), Theorem 3.9).

As we discuss in more detail in Section 3, the model given by (1), which we refer to as the *Hurdle model*, can be shown to be equivalent to a finite mixture model of singular Gaussian distributions. In light of the observed symmetry in the positive single cell expression levels, linking the modeling of zero-inflation with Gaussian parameters for nonzero observations is both natural and convenient. This said, it is an interesting topic for future work to develop more refined models of the continuous expression arising when genes are "on."

We will base statistical inference in the Hurdle model on so-called neighborhood selection, where the neighborhood of each gene is inferred via penalized regression methods (Meinshausen and Bühlmann (2006)). Neighborhood selection is a state-of-the-art method for estimation and inference in potentially high-dimensional graphical models; see the review in Section 3.4 of Drton and Maathuis (2017). The main challenge in our setting is determining how to calibrate signal in the binary versus the continuous part. We solve this problem using an *anisometric* group-lasso penalty (Section 4).

1.3. *Outline*. The remainder of the paper is structured as follows. Section 2 discusses the parameter targeted in single cell gene expression experiments, and why it is not accessible from traditional bulk experiments. Section 3 develops

the parametric Hurdle model for single cell gene expression, as specified in (1), and discusses conditional independence in this setting. Section 4 gives a detailed account of estimation of graphical models using neighborhood selection via penalized regression. Section 5 provides a simulation study that demonstrates the benefits of our approach. In Section 6, we analyze the aforementioned experiment on Tfh cells. Since the data set contains selected gene profiles that were available for both single- and several-cell aggregates, we are able to highlight the refined inferences that can be obtained from single cell data. In Section 7, we analyze data on mouse dendritic cells, which are of far higher dimensionality than the Tfh cell data. Our analyses show in particular that modeling the zero-inflation may uncover distinct networks compared to existing approaches. We conclude with a discussion in Section 8, where we highlight interesting problems for future research, in particular, in graphical modeling. A supplement, McDavid et al. (2019), contains expanded derivations and details on simulation scenarios.

2. Single cell versus bulk expression experiments. Protocols for bulk gene expression experiments, such as for Illumina TrueSeq, call for 100 nanograms of total mRNA, hence require hundreds to thousands of cells. On the one hand, this biological "summation" over many of cells is expected to yield sharper inference on the mean expression level of each gene. However, it can also be expected to distort any conditional (in-)dependences present between genes.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be i.i.d. random vectors taking values in $\mathbb{R}^m$, with $\mathbf{Y}_i$ representing the copy numbers of m transcripts present in the ith single cell. Now suppose the n cells are aggregated, and the total expression is measured using a linear quantification that reports values proportional to the input counts of mRNA. The expression observed in this *bulk* experiment is then

$$\mathbf{Z} \propto \sum_{i}^{n} \mathbf{Y}_i,$$

with the constant of proportionality typically a semi-empirical normalization factor, such as Transcripts Per Million or Fragments Per Kilobase Million. Although most bulk experiments are designed to test for differences in mean expression due to experimental treatments and lack extensive replication within a condition, *stochastic profiling* (Janes et al. (2010)) experiments have provided i.i.d. replicates of $\mathbf{Z}$ suitable for estimating higher order moments. However, when the distribution of $\mathbf{Y}_i$ obeys some conditional independence relationships, in general the distribution of $\mathbf{Z}$ does not obey these same relationships.

For example, take $m = 3$ and suppose that the $\mathbf{Y}_i$ are i.i.d. samples from a trivariate distribution supported on $\{0, 1\}^3$. Let $[Y_1, Y_2, Y_3]$ be a random vector following this distribution, and let $p_{ijk} = P(Y_1 = i, Y_2 = j, Y_3 = k)$ be the joint probabilities. Then Y_1 and Y_3 are conditionally independent given Y_2 (in symbols, $Y_1 \perp\!\!\!\perp Y_3 | Y_2$)

if and only if the two matrices $(p_{i0k})_{ik}$ and $(p_{i1k})_{ik}$ have rank 1 (Drton, Sturmfels and Sullivant (2009), Proposition 3.1.4). Yet even summing over only $n = 2$ cells, the random vector $\mathbf{Z} = \mathbf{Y}_1 + \mathbf{Y}_2 \equiv [Z_1, Z_2, Z_3]$ taking values in $\{0, 1, 2\}^3$ generally does not have $Z_1 \perp\!\!\!\perp Z_3 | Z_2$.

When the $\mathbf{Y}_i$ are multivariate Normal, the conditional independence structure is preserved under convolution. Unfortunately, for non-Gaussian distributions this does not generally hold. As noted in our Introduction, single cell gene expression is generally bimodal and zero-inflated, so not plausibly described by a multivariate Normal distribution. Therefore, even though for large enough n the distribution of the bulk experiment $\mathbf{Z}$ might approach multivariate (log-)normality, the networks estimated from graphical modeling of bulk data will not reflect conditional independences that hold among expression levels in single cells.

3. Hurdle models. Univariate Hurdle models arise from modification of a density through excision of points in the support and assignment of positive masses to these points. Targeting zero-inflation, our excision point is the origin. Let $v_y = I_{\{y \neq 0\}}$ be the indicator function for a nonzero value of the observation y. Then the Hurdle model derived from a Normal distribution with mean ξ and precision τ^2 has density

$$f(y) = \exp\{v_y[1/2\log(\tau^2/(2\pi)) + \log p/(1-p) - \xi^2\tau^2/2] + y\xi\tau^2 - y^2\tau^2/2 + \log(1-p)\} \tag{2}$$

with respect to the measure λ_0 that is the sum of the Lebesgue measure and a Dirac mass at zero. Here, $P(V_y = 1) = p \in (0, 1)$ is a mixing weight representing the chance of observing a nonzero value. Varying p, ξ and τ^2, one obtains an exponential family with sufficient statistic y, $-y^2/2$ and v_y, and associated natural parameters $h = \xi\tau^2$, $k = \tau^2$ and

$$g = 1/2\log(\tau^2/(2\pi)) + \log p/(1-p) - \xi^2\tau^2/2.$$

3.1. *Multivariate Hurdle models.* A plausible model for the joint distribution of a random vector $\mathbf{Y} = [Y_1, \ldots, Y_m]$ representing single cell gene expression puts positive mass on every one of the 2^m coordinate subspaces (Figure 2), including the origin when all genes are "off" and the entire space $\mathbb{R}^m$ when all genes are "on." Assigning positive mass to the coordinate subspaces generalizes the univariate construction from (2). As it is easiest to construct this model conditionally, we introduce the vector $\mathbf{V} = [V_1, \ldots, V_m]^T \equiv [I_{\{y_1 \neq 0\}}, \ldots, I_{\{y_m \neq 0\}}]^T$ that indicates the nonzero coordinates of $\mathbf{Y}$. Throughout, our notation suppresses the dependence of $\mathbf{V}$ on $\mathbf{Y}$. We emphasize that specification of the distribution of the multivariate Bernoulli random vector $\mathbf{V}$ simply amounts to specification of a 2^m probability table.

For any vector $\mathbf{v} = [v_1, \ldots, v_m] \in \{0, 1\}^m$, define the subspace $\mathbb{R}^{\mathbf{v}} = \prod_{i=1}^m \mathbb{R}^{v_i}$ where we set $\mathbb{R}^0 = \{0\}$. So, $\mathbb{R}^{\mathbf{v}}$ is the coordinate subspace corresponding to the

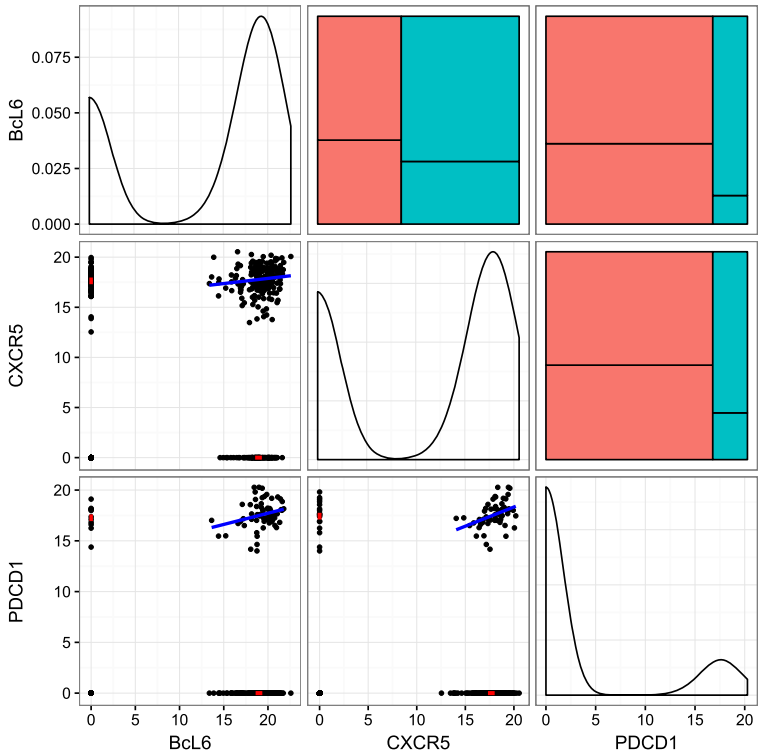

FIG. 2. *Scatter plots of inverse cycle threshold* (40-Ct) *measurements* $\mathbf{y}$ *from a quantitative PCR* (*qPCR*)*-based single cell gene expression experiment* (*lower panels*). *The cycle threshold* (*Ct*) *is the PCR cycle at which a predefined fluorescence threshold is crossed, so a larger inverse cycle threshold corresponds to greater log-expression* (*McDavid et al.* (2013)). *Measurements that failed to cross the threshold after* 40 *cycles are coded as* 0. *The upper panels show mosaic plots of each pair of contigency tables that can be formed from the indicator functions* $[\mathbf{v}_\mathbf{y}]_i = I_{\{y_i \neq 0\}}$. *On the lower panels, the linear regression on positive pairs of observations is indicated in blue, while the conditional mean values are indicated in red.*

nonzero entries of $\mathbf{v}$. Similarly, define $PD(\mathbf{v})$ to be the cone of $m \times m$ symmetric matrices that have nonzero entries only in rows and columns indexed by i with $v_i = 1$, and for which the submatrix given by these rows and columns is positive definite. Now suppose that the conditional distribution of $\mathbf{Y}$ given $\mathbf{V}$ is multivariate Normal and, specifically,

$$(\mathbf{Y}|\mathbf{V} = \mathbf{v}) \sim \mathcal{N}\big(\boldsymbol{\mu}(\mathbf{v}), \boldsymbol{\Sigma}(\mathbf{v})\big) \tag{3}$$

with mean vector $\boldsymbol{\mu}(\mathbf{v}) \in \mathbb{R}^\mathbf{v}$ and covariance matrix $\boldsymbol{\Sigma}(\mathbf{v}) \in PD(\mathbf{v})$. The normal distribution in (3) is singular (see Section 2 in supplement McDavid et al. (2019) for details) and supported on the subspace $\mathbb{R}^\mathbf{v}$.

In the applications, we have in mind the dimension m will be large enough so that it is infeasible to accurately estimate a general 2^m probability table for the

distribution of $\mathbf{V}$, and a collection of 2^m mean vectors and covariance matrices for the conditional distribution of $\mathbf{Y}$. We thus proceed to formulate a more parsimonious pairwise interaction model. While of far lower dimension, the pairwise model allows one to capture interesting conditional (in-)dependences.

First, we assume $\mathbf{V}$ to follow an Ising model with joint probabilities

$$p(\mathbf{v}) \equiv P(\mathbf{V} = \mathbf{v}) \propto \exp\bigl(\mathbf{v}^T \mathbf{G} \mathbf{v}\bigr), \qquad \mathbf{v} \in \{0, 1\}^m, \tag{4}$$

where $\mathbf{G}$ is a symmetric interaction matrix in $\mathbb{R}^{m \times m}$. Second, we assume that the conditional normal distribution of $\mathbf{Y}$ given $\mathbf{V} = \mathbf{v}$ has log-density

$$\log f(\mathbf{y}|\mathbf{V} = \mathbf{v}) = \mathbf{v}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C'(\mathbf{H}, \mathbf{K}), \qquad \mathbf{y} \in \mathbb{R}^{\mathbf{v}}, \tag{5}$$

with respect to Lebesgue measure restricted to the subspace $\mathbb{R}^{\mathbf{v}}$. In (5), $\mathbf{H}$ and $\mathbf{K}$ are two $m \times m$ interaction matrices that do not vary with $\mathbf{v}$, and $C'(\mathbf{H}, \mathbf{K})$ is a normalization constant. The matrix $\mathbf{K}$ is symmetric and positive definite, but $\mathbf{H}$ may be arbitrary from $\mathbb{R}^{m \times m}$. Putting the two pieces from (4) and (5) together, the joint density of $\mathbf{Y}$ with respect to the product measure λ_0^m simplifies to

$$f(\mathbf{y}) = \exp\left\{ \mathbf{v}^T \mathbf{G} \mathbf{v} + \mathbf{v}^T \mathbf{H} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{K} \mathbf{y} - C(\mathbf{G}, \mathbf{H}, \mathbf{K}) \right\}, \qquad \mathbf{y} \in \mathbb{R}^m. \tag{6}$$

We recognize an exponential family with three interaction matrices $\mathbf{G}$, $\mathbf{H}$ and $\mathbf{K}$ as natural parameters and the three statistics $\mathbf{v}\mathbf{v}^T$, $\mathbf{v}\mathbf{y}^T$ and $\mathbf{y}\mathbf{y}^T$ sufficient.

Let $\mathcal{I} \equiv \mathcal{I}(\mathbf{V})$ be the $m \times m$ diagonal matrix with (i, i) entry equal to V_i. Then for any vector $\mathbf{x} \in \mathbb{R}^m$ the product $\mathcal{I}\mathbf{x}$ is the vector that has the ith coordinate replaced by zero for all indices i with $Y_i = V_i = 0$. Similarly, multiplying $\mathcal{I}$ from left and right to a matrix zeros out all but the principal submatrix determined by this set of indices. Using this notation, the pairwise Hurdle model from (6) corresponds to the particular choice of

$$\boldsymbol{\mu}(\mathbf{v}) = (\mathcal{I}\mathbf{K}\mathcal{I})^{-}\mathbf{H}\mathbf{v}, \qquad \boldsymbol{\Sigma}(\mathbf{v}) = (\mathcal{I}\mathbf{K}\mathcal{I})^{-} \tag{7}$$

for the mean vectors and covariance matrices in the conditional specification from (5). In (7), A^- denotes the Moore–Penrose pseudo-inverse of a matrix A. From the perspective of (7), the pairwise Hurdle model is a mixture of 2^m singular Gaussian distributions whose mean vectors and covariance matrices are derived from one precision matrix $\mathbf{K}$ and an interaction matrix $\mathbf{H}$.

The notation we used in the conditional specification of the multivariate Hurdle model follows Lauritzen (1996), who describes conditional Gaussian (CG) models with *inhomogeneous, nonsingular* precision $\mathbf{K}(\mathbf{v})$ that can depend on the discrete set of covariates in arbitrary, positive-definite fashion. These models have been considered more recently by Lee and Hastie (2013) and Cheng et al. (2017). Our formulation differs from the traditional CG models by involving singular distributions with means and covariance matrices that exhibit structured inhomogeneity.

3.2. *Conditional distributions identify interaction parameters.* The normalizing constant C in equation (6) is a difficult to compute sum of 2^m terms. This is expected as already the distributions in the Ising model from (4) have an intractable normalization constant for moderately large m. Fortunately, the univariate full conditional distributions obtained from (6) have tractable normalizing constants and identify the parameters from a given row/column of the interaction matrices $\mathbf{G} = (g_{ab})$, $\mathbf{H} = (h_{ab})$ and $\mathbf{K} = (k_{ab})$.

Fix a coordinate b, and define its complement $A = \{1, \ldots, m\} \setminus \{b\}$. Consider now the density $f(\mathbf{y})$ from (6) as a function of only y_b, that is, $\mathbf{y}_A = [y_i : i \in A]$ is fixed, and write $f_{[b|A]}$ for the conditional density of y_b given $\mathbf{y}_A$. Then noting that $v_i y_i = y_i$ and $v_i^2 = v_i$, we have

$$\log f_{[b|A]}(\mathbf{y}) = v_b g_{[b|A]} + y_b h_{[b|A]} - \frac{1}{2} y_b^2 k_{[b|A]} - C_{[b|A]}, \qquad y_b \in \mathbb{R}, \tag{8}$$

where $C_{[b|A]}$ does not depend on y_b and

$$g_{[b|A]} = g_{bb} + 2\mathbf{g}_{bA}\mathbf{v}_A + \mathbf{h}_{bA}\mathbf{y}_A, \tag{9}$$

$$h_{[b|A]} = h_{bb} + \mathbf{h}_{Ab}^T\mathbf{v}_A - \mathbf{k}_{bA}\mathbf{y}_A, \tag{10}$$

$$k_{[b|A]} = k_{bb}.$$

The conditional density $f_{[b|A]}$ is thus a univariate Hurdle density as specified in (2) with natural parameters $g_{[b|A]}$, $h_{[b|A]}$ and $k_{[b|A]}$.

The three natural parameters are obtained from linear predictors that depend on a design matrix constructed from $\mathbf{y}_A$ and $\mathbf{v}_A$. For example, we may write

$$g_{[b|A]} = g_{bb} + \sum_{a \in A} X_a \begin{bmatrix} g_{ba} \\ h_{ba} \end{bmatrix}$$

for $X_a = [v_a, y_a]$. The linear predictor for $h_{[b|A]}$ can be written analogously. We note that if the data include additional nuisance covariates $\mathbf{W}_0$ that describe each experimental unit then these can be included by augmenting the linear predictor to

$$g_{[b|A]} = \mathbf{W}_0^T \mathbf{g}_{b0} + g_{bb} + \sum_{a \in A} X_a \begin{bmatrix} g_{ba} \\ h_{ba} \end{bmatrix} \tag{11}$$

with $\mathbf{g}_{b0}$ being the parameters capturing the effects of the covariates. From this perspective, the conditional distribution in (8) defines a vector generalized linear model, parametrized by three natural parameters $g_{[b|A]}$, $h_{[b|A]}$ and $k_{[b|A]}$, the first two of which are modeled as a linear function of the expression of other genes.

3.3. *Conditional independence graphs.* The dependence structure of the random vector $\mathbf{Y} = [Y_1, \ldots, Y_m]$ may be summarized in its *conditional independence graph*. This is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, \ldots, m\}$ and an edge set $\mathcal{E}$ that is determined by the conditional independences in $\mathbf{Y}$. More precisely, the edges in $\mathcal{E}$ are those two-element sets $\{a, b\} \subset \mathcal{V}$ for which Y_a and Y_b

are conditionally dependent given the remaining variables, that is, $\mathbf{Y}_{\mathcal{V}\setminus\{a,b\}}$. In our case, $\mathbf{Y}$ has a density f as in (6). The dominating measure is a product measure, and f is positive and continuous. Hence, the Hammersley–Clifford theorem assures that the conditional independence graph of $\mathbf{Y}$ has an edge $\{a, b\}$ if and only if the four possible ab interactions are zero, so

$$g_{ab} = h_{ab} = h_{ba} = k_{ab} = 0; \tag{12}$$

see Lauritzen (1996), Chapter 3. This fact is also evident from the form of the conditional distributions detailed in (8), (9) and (10). It motivates the neighborhood selection procedure developed in the next section.

4. Neighborhood estimation via penalized regression. In the single cell experiments to which we envision applying this method, the number of cell replicates, n, is larger than the sample sizes seen in typical bulk mRNA experiments. However, it is still often the case that the number of genes m is larger than the number of cell replicates. We are thus in a setting that benefits from application of methods from "high-dimensional statistics," though emerging technologies are increasing available sample sizes.

4.1. *Related work.* Under scenarios in which $n, m \to \infty$ while satisfying that $n > Cd^{\phi}(\log m)^{\psi}$, where C, ϕ and ψ are constants that depend on the model and d is the maximum vertex degree of the conditional independence graph, penalized regression has been shown to consistently identify the graph of multivariate Normal models (Meinshausen and Bühlmann (2006)), of Ising (auto-logistic) models (Ravikumar, Wainwright and Lafferty (2010)) and of exponential family graphical models (Chen, Witten and Shojaie (2015), Yang et al. (2014)). While this paper was in preparation, Tansey and Hernan Madrid Padilla (2015) further extended this line of work to general *vector space graphical models* that include the multivariate Hurdle model as a special case. However, the standard (isometric) group-lasso they propose for estimation of the conditional independence graph does not account for heterogeneity in the scaling of predictors in the conditional distributions. The anisometric group-lasso we propose in the following section yields drastic improvements in finite samples.

4.2. *Anisometric penalty.* Throughout this section, we fix an index b and consider the conditional distribution Y_b given the other variables in $\mathbf{Y}_A$ for $A = \{1, \ldots, m\} \setminus \{b\}$. For any $a \in A$, define the parameter vector $\theta_a = [g_{ba}, h_{ba}, h_{ab}, k_{ba}]$. By (12), $Y_b \perp\!\!\!\perp Y_a | \mathbf{Y}_{A\setminus\{a\}}$ if and only if $\theta_a = 0$.

Let $\theta = [\theta_a : a \in A]$, and let

$$P_\lambda(\theta) = \lambda \sum_{a\in A} \sqrt{\theta_a^T \theta_a} \tag{13}$$

be the group lasso penalty for tuning parameter $\lambda \geq 0$. Maximization of the penalized conditional log-likelihood function

$$\log f_{[b|A]}(\mathbf{y}) - P_\lambda(\theta)$$

can lead to a solution that is sparse in parameter blocks, that is, some of the subvectors θ_a are zero. The penalty is equivalent to placing a sequence of independent, multivariate Laplace priors on blocks of θ and reporting the MAP (Eltoft, Kim and Lee (2006)).

Viewed as a prior, the standard group-lasso penalty from (13) implicitly assumes that each variable in each block has a similar effect size. This may be reasonable if the variables in each block are measured in comparable units, but is problematic otherwise. For example, if covariate X_1 is measured in meters, while covariate X_2 in centimeters, then the distribution of effect sizes for X_2 would be 100-times more dispersed than the distribution of effect sizes for X_1. In penalized GLMs, this is typically enforced "at run time" by ensuring covariates are on comparable scales, or Z-scoring each column of the design matrix if no intrinsic scale exists.

In our setting of a vector regression, terms from linear predictor $g_{[b|A]}$ and linear predictor $h_{[b|A]}$ end up together in blocks, and these coefficients are not necessarily comparable, as one specifies log-odds of $E(V_b|V_A = 0)$ while the other specifies conditional expectations of $E(Y_b|Y_A)$. Rescaling does not resolve this, since the same design matrix $X_a = [V_a, Y_a]$ is used in each linear predictor, and in any case, rescaling generally alters the solution (Simon and Tibshirani (2012)). Instead, we propose replacing the isometric ℓ_2 norm in the sum in (13) so that the penalty is

$$P_{\mathbf{H},\lambda}(\theta) = \lambda \sum_{a \in A} \sqrt{\theta_a^T \mathbf{H}_{aa} \theta_a}. \tag{14}$$

Here, $\mathbf{H} \equiv \operatorname{diag}(\mathbf{H}_{aa})$ is a block-diagonal, positive-definite matrix that allows terms from the linear predictors to have different scales of penalty. It also accounts for correlation between components of θ_a, since columns of the design are correlated due to both v_a and y_a appearing as predictors.

If prior information existed, the matrix $\mathbf{H}$ could be chosen accordingly, with interpretation as a multivariate Laplace prior. Absent prior information, setting $\mathbf{H}$ equal to the Fisher information under a null model $\theta_a = 0$ for all a results in variable selection approximately equal to conducting score tests, with exact equivalence holding under a null hypothesis of $\theta_a = 0$ for all a; see Proposition 1 in McDavid et al. (2019).

4.3. *Computation*. In Algorithm 1, we outline the proposed neighborhood selection, allowing for possible nuisance covariates $\mathbf{W}$. The nuisance covariates $\mathbf{W}$ might just be an intercept column, but generally could be any cell-level covariate deemed relevant. The smooth and concave function in line 7 can be maximized using any Newton-like algorithm (e.g., BFGS). The objective in line 10 is a sum of a

Data: Expression matrix $\mathbf{Y} \in \mathbb{R}^{n\times m}$, nuisance covariates $\mathbf{W} \in \mathbb{R}^{n\times q}$, penalty path $\mathbf{\Lambda}$.
Working parameters: Unpenalized nuisance parameters $\theta_0 \in \mathbb{R}^{2q+1}$, edge parameters $\theta \in \mathbb{R}^{4(m-1)}$.
Result: Neighborhoods $ne(i, \lambda)$, $1 \le i \le m$, $\lambda \in \mathbf{\Lambda}$

for $b \in \{1, \ldots, m\}$ **do**
  $A \leftarrow \{1, \ldots, m\} \setminus \{b\}$;
  $\mathbf{X} \leftarrow [\mathbf{W}, \mathbf{Y}_A, \mathbf{V}_A]$;
  $\theta_0 \leftarrow [g_{bb}, \mathbf{g}_{b0}^T, h_{bb}, \mathbf{h}_{b0}^T, k_{bb}]$;
  $\theta \leftarrow [\mathbf{g}_{bA}, \mathbf{h}_{bA}, \mathbf{h}_{Ab}^T, \mathbf{k}_{bA}]$;
  Let $\log f_{[b|A]}(\theta_0, \theta)$ return the log-density (8) evaluated at $[\theta_0, \theta]$ with covariate matrix $\mathbf{X}$.
  $\bar{\theta}_0 \leftarrow \operatorname{argmax}_{\theta_0} \log f_{[b|A]}(\theta_0, \theta = 0)$;
  $\mathbf{H} \leftarrow \nabla^2 \log f_{[b|A]}(\bar{\theta}_0, 0)$;
  for $\lambda \in \mathbf{\Lambda}$ **do**
    $[\hat{\theta}_0, \hat{\theta}] \leftarrow \operatorname{argmax}_{\theta_0,\theta} \log f_{[b|A]}(\theta_0, \theta) - P_{\mathbf{H},\lambda}(\theta)$;
    Let $ne(b, \lambda)$ contain vertex a whenever any of $\hat{\mathbf{g}}_{bA}, \hat{\mathbf{h}}_{Ab}, \hat{\mathbf{h}}_{bA}, \hat{\mathbf{k}}_{Ab} \neq 0$.
  end
end

Algorithm 1: Neighborhood selection

concave, smooth function and a structured concave function and can be efficiently solved using proximal gradient ascent (Parikh and Boyd (2014)). In particular, one may exploit the fact that although the proximal operator

$$\underset{\gamma}{\text{prox}}(x) = \underset{u}{\text{argmax}} \frac{1}{\gamma}\|x - u\|_2^2 + \sum_{a\in A} \sqrt{u_a^T H_{aa} u_a}$$

is not available in the familiar form of a soft-thresholding operator as in the isometric group-lasso, the proximal operator of the *anisometric* group-lasso can be efficiently found via a line search after one-time precalculation of the singular value decomposition of H_{aa} (Foygel and Drton (2010)). Throughout the inner-loop, warm starts are exploited for $\hat{\theta}$ as λ varies. Active set heuristics using the strong rules of Tibshirani et al. (2012) yield computational gains for sparse solutions with large m. The algorithm yields, for each node, a sequence of neighborhoods over a sequence of tuning parameters $\mathbf{\Lambda}$. These neighborhoods need not be consistent, in the sense that for some element of $\mathbf{\Lambda}$ it could be that $b \in \text{Ne}(a)$ but $a \notin \text{Ne}(b)$. We resolve that by adopting an "or" rule. In the accompanying software,[4] the algorithm is written in a combination of `R` and `C++`. Timings for the

[4] Available at https://github.com/amcdavid/HurdleNormal.

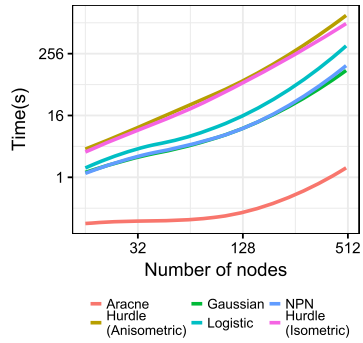

FIG. 3. *Average timings for graph estimation algorithms as a function of the number of nodes.*

proposed method and competitors (described further in Section 5) are shown in Figure 3.

5. Simulations. We consider a series of simulations under several sets of underlying (i) graph topologies, (ii) parametric models, (iii) sample sizes and (iv) number of vertices. We summarize the considered setups here and defer details to Section 3 in the Supplementary Material (McDavid et al. (2019)). The number of observations n varies from 100 to 12,500. In the *chain* graph topology, the number of vertices varies from $m = 16$ to $m = 128$, while in the *e. coli* graph topology, $m = 500$. The parametric models include the pairwise hurdle model (6), the hurdle model under contamination by t_8 noise, a logistic/Ising model and a Gaussian/logistic censoring model specified in Supplementary Table 1. The pairwise hurdle model is said to be *complete* if for each edge present in the graph, all of the corresponding entries in each of the three interaction matrices are nonzero. The pairwise hurdle model is said to be *G-minimal* when H and K are diagonal matrices and only G contains nonzero off-diagonal entries. In this case, the G-minimal model is equivalent to a logistic/Ising model.

5.1. *Methods compared and default tunings.* Six methods were examined to test graph structure inference, and are described in Supplemental Table 1. The Hurdle models are fit using the accompanying software `HurdleNormal` version 0.98.2, while the Logistic, Gaussian and NPN models are fit using the `R` package `glmnet` version 2.0-5 (via the `autoGLM` function in `HurdleNormal`). The Aracne method is fit using package `netbenchmark` version 1.6.0. For methods 1–5, neighborhoods are stitched together using an "or" rule, that is, vertices a and b are adjacent if either $b \in \text{ne}(a)$ or $a \in \text{ne}(b)$.

In Figure 4, various fixed tunings are shown. In the *oracle* tuning, the graph with maximum sensitivity subject to FDR $< 10\%$ is shown. This tuning is not available in practice, but shows the maximum achievable performance of each

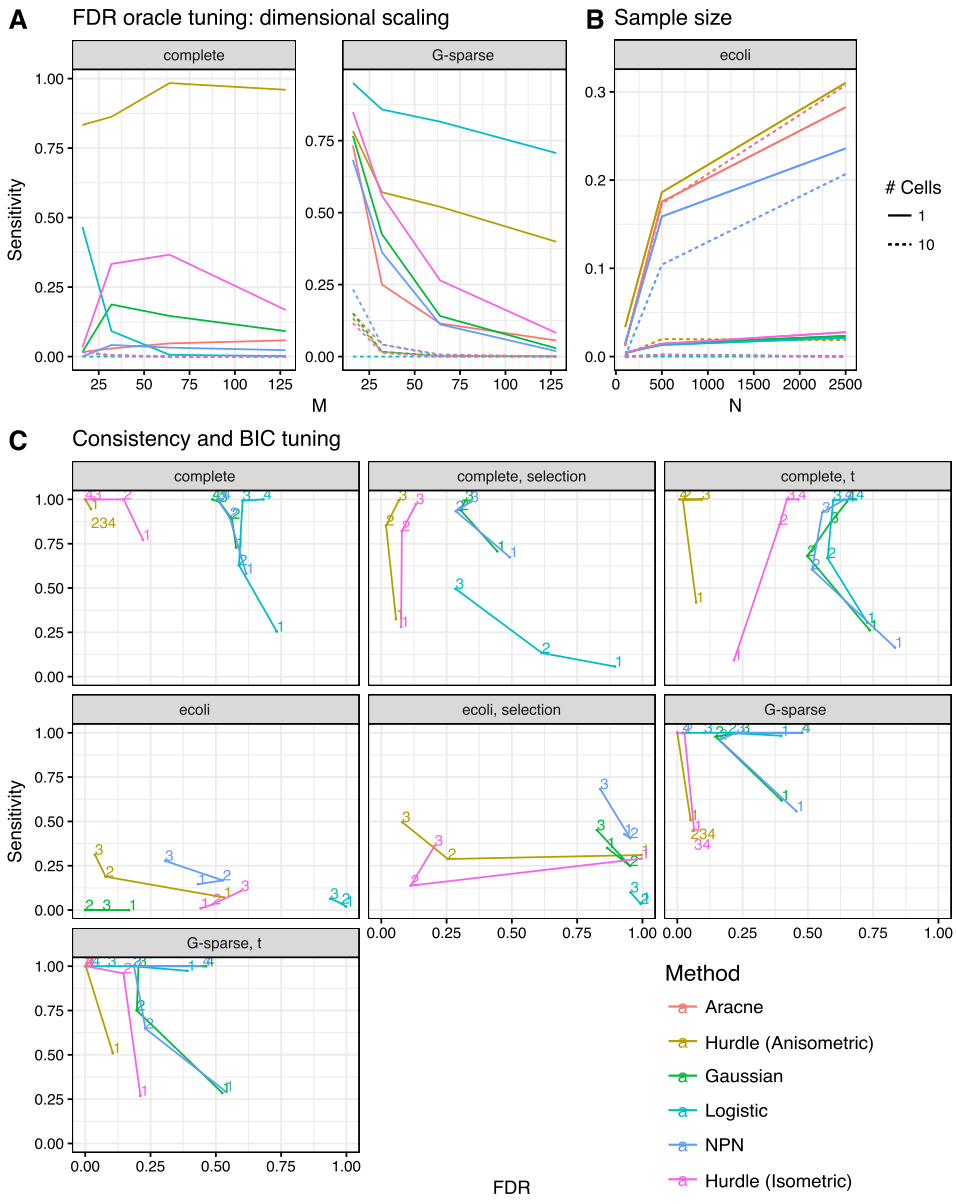

FIG. 4. *Dimensional* (*A*) *and sample size scaling* (*B*) *of six different network inference algorithms applied to simulated data under oracle FDR tuning. Data are generated from the multivariate hurdle model* (6) *under chain graphs* (*A*) *and* e. coli *graph* (*B*). *Panel* (*C*) *shows network selection consistency of various methods using the Bayesian Information Criterion under various models described in the Supplementaey Material. The paths trace out the changes in FDR and sensitivity as the sample size increases geometrically from* 100 (1) *to* 12,500 (4).

method. With the *BIC* tuning, we employ the Bayesian Information Criteria on the pseudo-likelihood

$$\mathrm{BIC}_\lambda = \sum_{b \in \mathcal{V}} -2 \log f_{[b|A]}(\hat{\theta}_{b,\lambda}) + \|\theta_{b,\lambda}\|_0 \log n,$$

where $\theta_{b,\lambda}$ is the penalized solution at penalty λ for vertex b, $\|\theta_{b,\lambda}\|_0$ is the number of nonzero entries, and $\hat{\theta}_{b,\lambda}$ is the (unpenalized) maximum pseudo-likelihood estimate for the nonzero entries. The BIC solution is the one that minimizes BIC_λ. This tuning is available for methods 1–5. In the case of the the Aracne method, the BIC is unavailable as no likelihood is defined.

5.2. *Results*. Thirty simulation replicates sufficed to bound the simulation-induced Monte Carlo standard error of the mean $< 5 \times 10^{-3}$ for FDR and < 0.02 for the sensitivity.

The simulations show that misspecified estimation procedures perform poorly when model (6) is the data generating distribution. When a FDR-controlling oracle is available, the anisometric Hurdle model can dominate other methods in edge-sensitivity (Figure 4A–B). However, when the Hurdle model is over-parameterized as in the G-sparse scenarios, the minimal Logistic model is superior, though the anisometric ℓ_1 penalty partially ameliorates this gap. In very simple chain-graph scenarios, it is nigh-impossible to recover a network using 10-cell data. The *e. coli* network provides a counterexample where 10-cell data nearly equals the performance available from single cell data. This may be due to the hub-and-spoke nature of the *e. coli* network, so the effect of *marginalization by convolution* tends to only add more connections between the hub and its neighborhood. The *e. coli* data and chain-graphs suggest that collecting single cell data, and estimating graph structure with a method that accommodates zero inflation can accurately discover a wide variety of network topologies.

More seriously, ignoring zero-inflation confounds use of information criteria to tune network size (panel C). On the other hand, the Hurdle model is robust to a variety of model departures, including contamination with t_8-distributed errors (labeled with "t"), and data generation under a Gaussian-Logistic censoring model. When the full solution path is examined (Figure 5), a practitioner who reported only the top few edges would often suffer from a large number of false positives when using methods not designed for zero-inflated data. For example, with $n = 100$ in the *e. coli* network, all methods, aside from the Hurdle have FDR exceeding 20%. The simulations also suggest that perfect recovery of gene networks is impractical at realistic sample sizes, even with a correctly specified model, motivating a form of meta-analysis on estimated graphs, discussed further in Section 7.2.

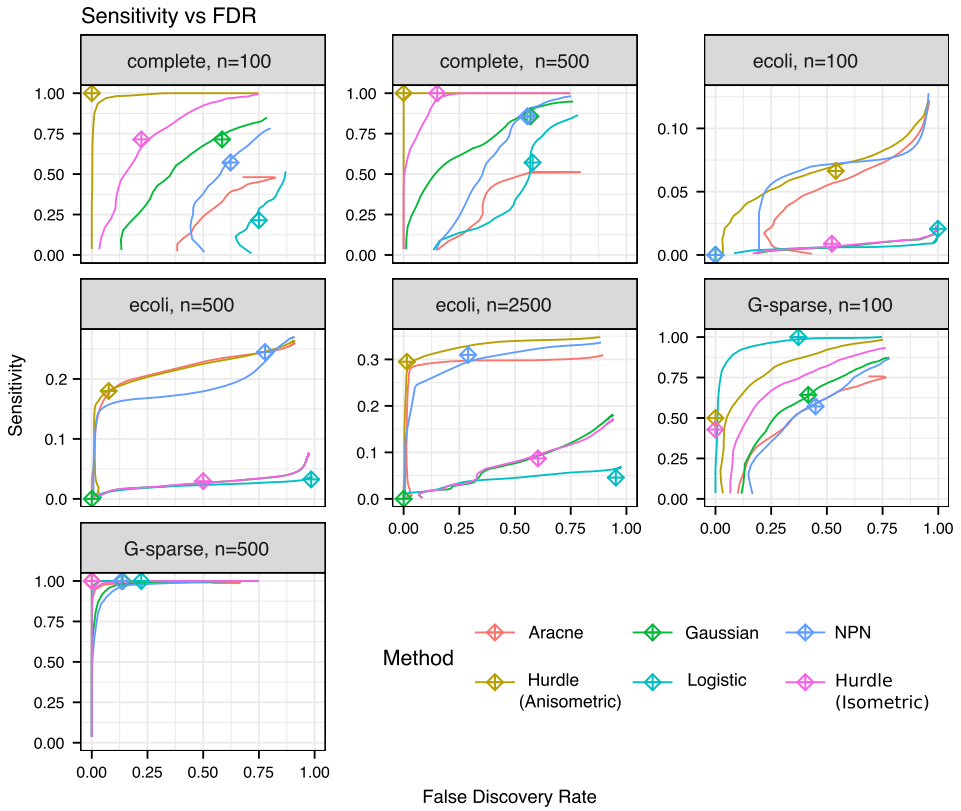

FIG. 5. *Sensitivity vs. FDR for solution paths from methods and scenarios described in Supplemental Table* 1. *The* $\oplus$ *symbol indicates the tuning selected by BIC.*

6. T follicular helper cells. Our simulations show that depending on the data generating scenario, the Hurdle method may substantially out-perform, or at least mimic the performance of other candidate methods. We next sought to see if methods would tend toward consensus in biologically derived single cell and 10-cell data, or if it were possible that the Hurdle method might offer unique insights. We considered co-expression networks in Tfh cells measured in eight healthy donors. 65 genes were selected for profiling via qPCR on the basis of their role in Tfh signaling and differentiation, generally with sparse expression across single cells (overall probability of expression 27%). 465 single cell, and 187 10-cell replicates were taken.

Figure 6 shows networks of approximately 24 edges estimated using Hurdle, Gaussian (with centered data, see Section 1 McDavid et al. (2019) and Logistic, and Gaussian model using 10-cell aggregates. The size of the network is a compromise between stability selected (Shah and Samworth (2013)) sizes of each procedure, which varies from 11 edges (Hurdle) to 32 edges (Gaussian).

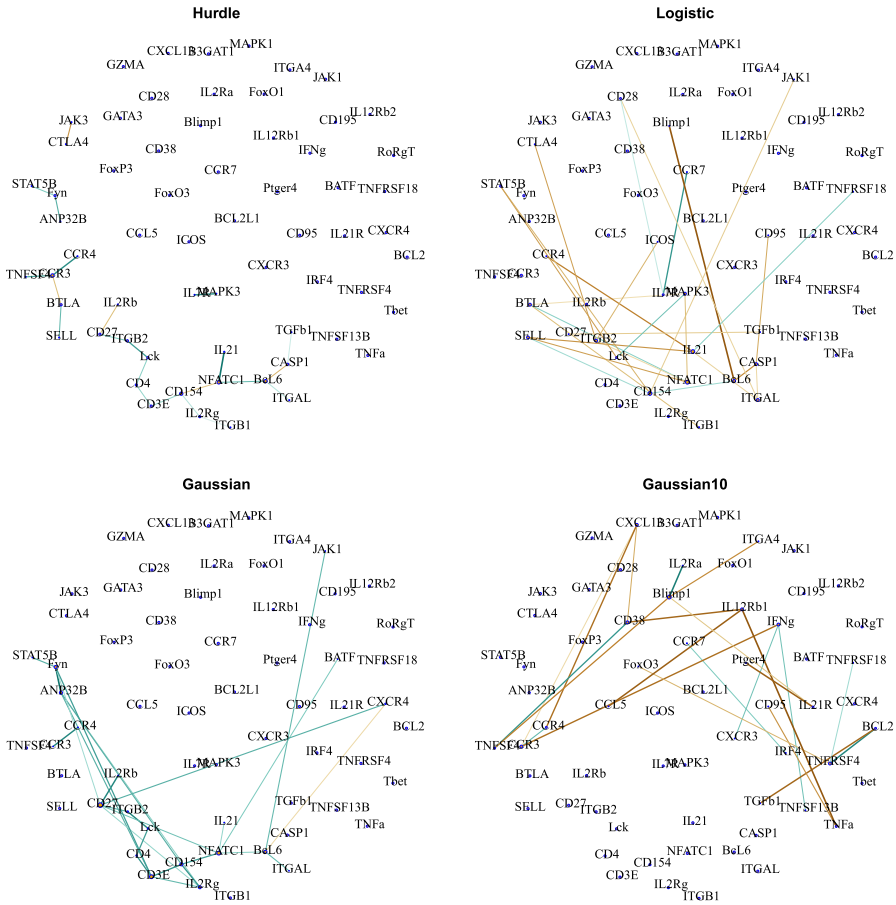

FIG. 6. *Networks of* 22 *edges estimated through neighborhood selection under the Aracne*, *Hurdle*, *logistic*, *Gaussian model* (*single cells*) *and Gaussian model* (10 *cell aggregates*) *in T follicular helper cells. Brown hues indicate estimated negative dependences, while blue-green hues indicate positive dependences. The edge width and saturation are larger for stronger estimated dependences.*

Normalized Hamming distances between the four methods, the Aracne method and the Gaussian model fit on the "raw," uncentered data are reported in Table 1. The Hurdle and Gaussian models are most similar, while the logistic and Gaussian 10-cell network are quite distinct. The Gaussian(raw) model on untransformed data is similar to the logistic model, as distance of nonzero expression values from the origin is large compared to the variation among the nonzero values.

In the Hurdle network, the transcription factors NFATC1 (Nuclear factor of activated T-cells) and BCL6, and the signaling molecule CD154 and chemokine receptor CCR3 are hubs. NFATC1 has been found to promote transcription of cytokines IL21 (Hermann-Kleiter and Baier (2010)) and signaling molecule CD154 (Lan et al. (2005)), while BCL6 serves as a transcriptional repressor, and is one of the

TABLE 1

Dissimilarities ($\frac{\textit{Hamming Distance}}{\textit{Number of edges}}$) *between networks of size* 24 *estimated through various methods. The Gaussian*(10) *model is a Gaussian model estimated on* 10*-cell replicates, while the Gaussian*(*raw*) *data is estimated on single cells without centering the data. The remaining models are described in Section* 5

	Gaussian(10)	Gaussian	Gaussian(raw)	Hurdle	logistic
Aracne	1.00	0.92	0.92	1.00	1.00
Gaussian(10)		1.00	1.00	1.00	1.00
Gaussian			0.92	0.65	1.00
Gaussian(raw)				1.00	0.39
Hurdle					1.00

canonical markers constitutively expressed in Tfh cells. CTLA4 which has been described to inhibit inflammation, interacts negatively with inflammatory activator JAK3. The disconnected component of CCR3-CCR4-BTLA-SELL-TNFSF4 may hint at plasticity between Tfh cells and the related T-cell lineages Th1 and Th2. CCR3 and CCR4 are canonical markers of Th2 cells, while TNFSF4 (coding for OX40L) promotes Th2 development (de Jong et al. (2002)). Thus co-expression of these genes may suggest cells transitioning between Tfh and Th1 or Th2 states.

In the Gaussian network, though NFATC1, BCL6 and CD154 remain highly connected, CD27 now has highest degree and serves as a hub to receptors CXCR4, IL2Rb, IL2Rg, as well as ITGB2, NFATC1 and FYN. CD3e, the backbone responsible for transducing the T-cell receptor signal is connected with co-receptor CD4, CD154, IL2Rg, Fyn and ANP32B. The negative interactions between BTLA and CTLA4 are absent.

The logistic network consists primarily of negative interactions. The strongly negative BCL6–BLIMP1 edge is consistent with previously described antagonism between these genes (Johnston et al. (2009)). Interestingly, this edge is absent in the other networks.

7. Mouse dendritic cells. Shalek et al. (2014) exposed bone marrow-derived dendritic cells, from *mus musculus*, to lipopolysaccharide (LPS). LPS is a toxic compound secreted and structurally utilized by gram-negative bacteria and induces a cascade of changes in a cell's expression profile through several pathways. Cells were sampled after 0, 1, 2, 4 and 6 hours post-exposure. We estimated transcription networks using 4431 transcripts expressed in at least 20% of 65 cells sampled 2 hours after LPS exposure, at which interval transcription is expected to be undergoing a variety of dynamic changes. Rather than attempting to perform model selection on this limited sample size, we consider highly sparse ($< 0.01\%$ sparsity) networks of 700 edges, chosen to provide tractable visualization and illustration of the method. The BIC tunings (discussed subsequently) are decidedly larger.

7.1. *Selected networks*. In a Gaussian model, the network is star-shaped, with Mx1, Ccl17, Tax1bp3 and Ccl3 as hubs all with degrees ≥ 15, though none are directly interconnected (Figure 7). In all, 2.5% of nonisolated vertices contribute 50% of the edges in the network. With the exception of Tax1bp3, these hub genes are all immune-signaling related.

In the Hurdle model (Figure 8), the graph is more chain-like, with maximum degree 12: 7% of nodes provide 50% of the edges. The strongest hub, Mgl2 (also known as Cd301b), has been recently described to be involved in uptake and presentation of glycosylated antigens, such as LPS, by dendritic cells (Denda-Nagai et al. (2010)). A subconnected set of genes coding for MHC-II antigen presentation (H2ab1, H2eb1, H2aa) is the densest sub-component, and interconnected to

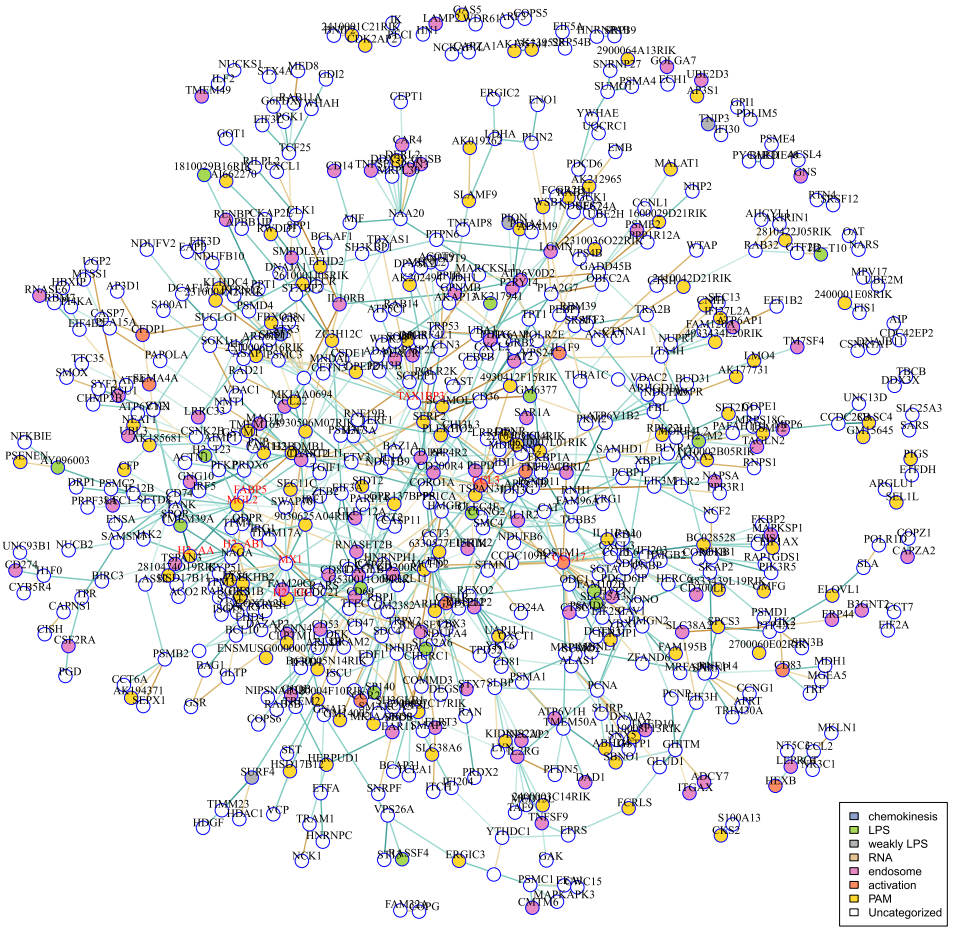

FIG. 7. *Core Gaussian model networks in LPS-treated mouse dendritic cells. Hub genes are shown in red. Vertex colors indicate gene ontology membership. Disconnected subgraphs with two vertices are suppressed.*

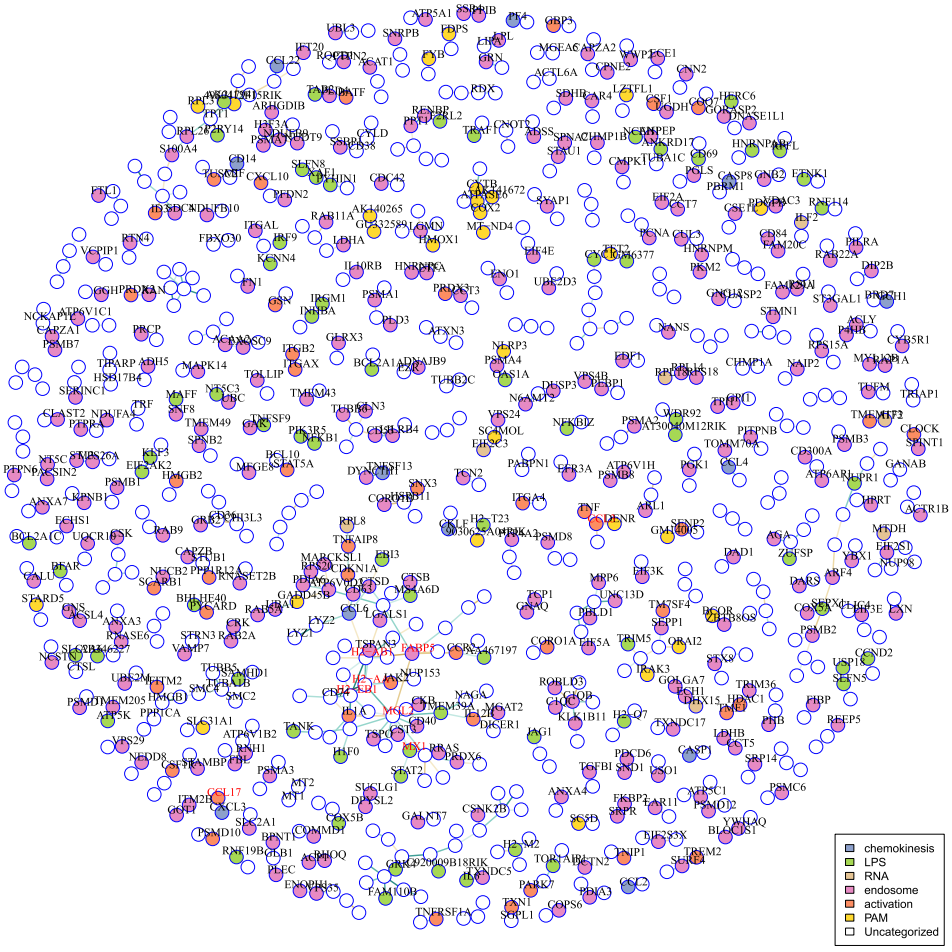

FIG. 8. *Core Hurdle model networks estimated in LPS-treated mouse dendritic cells. Hub genes are shown in red. Vertex colors indicate gene ontology membership. Disconnected subgraphs with two vertices are suppressed.*

Mgl2 as well as Fabp5. Increased expression of Fabp5 has been shown to increase expression of cytokines Il7 and Il18, hence is also involved in immune cell stimulation (Adachi et al. (2012)). Many of the neighbors of Mgl2, H2ab1, H2eb1, H2aa and Fabp5 are neighbors of the hub genes in the Gaussian graph, whereas Mx1, Ccl17 and Ccl3 are sparsely connected in the Hurdle network. Tax1bp3 is absent.

Using BIC, both the Gaussian and Logistic models yield networks with more than 25,000 edges, while the Hurdle selects a network of roughly 12,000 edges. The additional flexibility available in the Hurdle for modeling inter-node relationships may permit sparser graphs to describe the conditional dependence relationships. We also observe that the Hurdle synthesizes signal from both Gaussian and

Logistic networks. For sufficiently rich network sizes, the Gaussian and Hurdle and Logistic and Hurdle networks share 21% and 1% of possible edges, respectively, compared to only 0.08% of possible edges between the Gaussian and Logistic networks (binomial test $p < 10^{-6}$).

7.2. *Graphical geneset edge enrichment*. We consider how well the 700 edge networks recapitulate known relationships between genes using previously described functional annotations. The Gene Ontology Consortium (2015) provides a database of categories to which genes may be annotated if experimentally or computationally they are involved in a biological process. We note that networks may exhibit *intraconnection* within GO categories, and that some pairs of categories may exhibit preferential *interconnection*.

Each pair (i, j) of GO categories—including self-pairs—induces a *coloring* of vertices, coloring the vertices belonging to category i color c_i and category j color c_j. Vertices that do not belong to either i or j remain uncolored. Iterating through the $3987^2/2$ pairs of categories, we test for edge enrichment between colors. Suppose in the inferred graph of 700 edges, n_{ij} edges connect c_i-colored vertices to c_j vertices. If the colored vertices were completely connected with n_i vertices of color c_i and n_j vertices of color c_j, then there would be $m_{ij} = n_i \times n_j$ edges among them (with the obvious adjustment made for self-edges when $i = j$). Figure 9 depicts the procedure on four nodes. We now define an enrichment statistic as the hypergeometric tail probability

$$t_{ij} = P(N_{ij} > n_{ij}; 700, m_{ij}, 4431 \times 4430/2),$$

which is the probability of drawing n_{ij} colored balls, given 700 draws from urn containing $4431 \times 4430/2$ balls of which m_{ij} are colored.

This results in nearly 16 million enrichment statistics on the pairs of categories, which follow a complicated dependence structure under the series of null hypotheses that the observed edges being connected independent of coloring. The top 200 (smallest in magnitude) enrichment statistics $t_{(k)}$, $k < 200$ are compared to their distribution $P(t^*)$ under a Erdős–Rényi random graph model, yielding a Monte

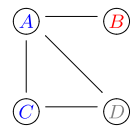

FIG. 9. *Overview of geneset edge enrichment analysis*. 1. *Vertices A and C belong to the blue category, while vertex B belongs to the red category. Vertex D belongs to neither.* 2. *There is* $n_{ij} = 1$ *blue-red intra-connection, while* $m_{ij} = 2$ *are possible given the* 4 *edges*. 3. *The enrichment statistic is the hypergeometric tail probability* $t_{ij} = P(N = 2; 4, 2, 6) = \frac{\binom{2}{2}\binom{4}{2}}{\binom{6}{4}} = 0.4$. 4. *The significance of the blue-red enrichment statistic would be ascertained by sampling from the null Erdos–Renyi model over all possible pairs of categories*.

Carlo p-value for each order statistic. A pair of colors (i, j) with rank $r_{ij} < 200$ is declared significant if $P(t_{ij} < t^*_{(r_{ij})}) < 0.05$ and $P(t_{(r)} < t^*_{(r)}) < 0.05$ for all $r < r_{ij}$, that is, it is significant at 5% and all smaller order statistics are also significant.

7.2.1. *Hurdle graphs tend to include intra-category enrichment.* In the Gaussian model, more than 100 pairs of categories (colors) are significantly enriched at an FDR of less than 10%, however in these pairs, only 6 correspond to intracategory enrichment (Figure 10). These are: response to salt stress, potassium channel regulator activity, extracellular exosome and three genesets containing genes with significant time-course differential expression in the original experiment. In the Hurdle model (Figure 11), 13 of 57 significantly enriched pairs form intra-

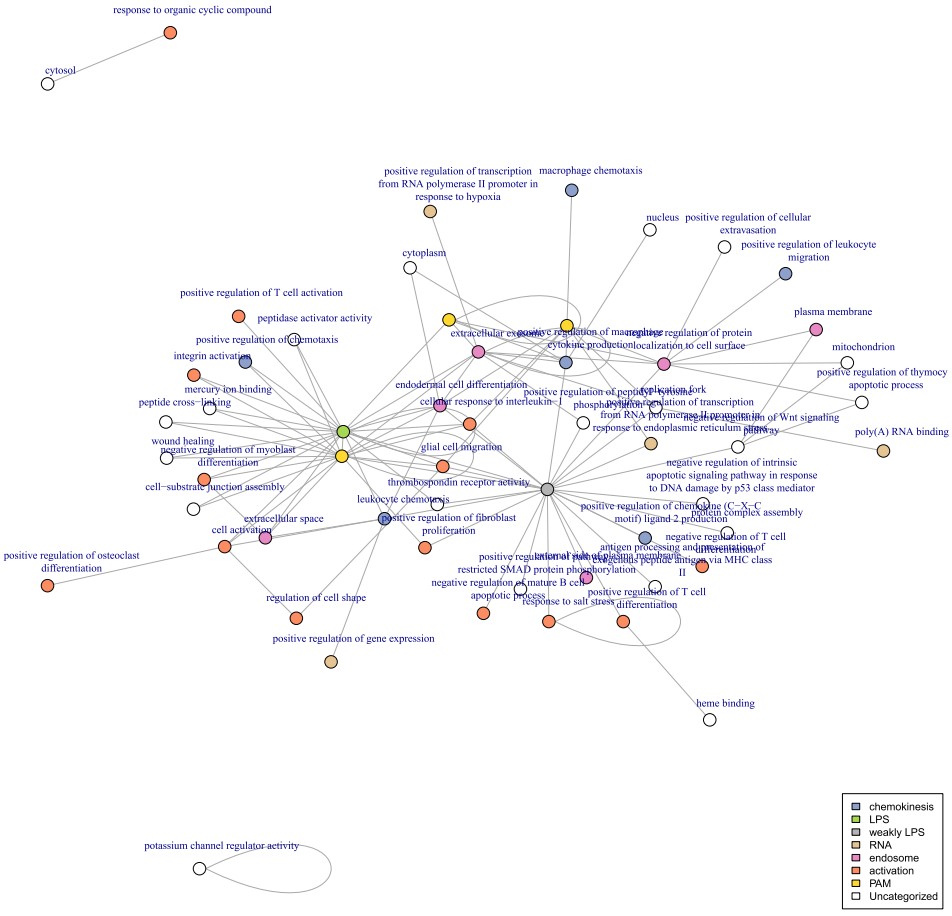

FIG. 10. *Modules enriched at FDR ≤ 10% using graphical geneset edge enrichment in mouse dendritic cells under Gaussian model.*

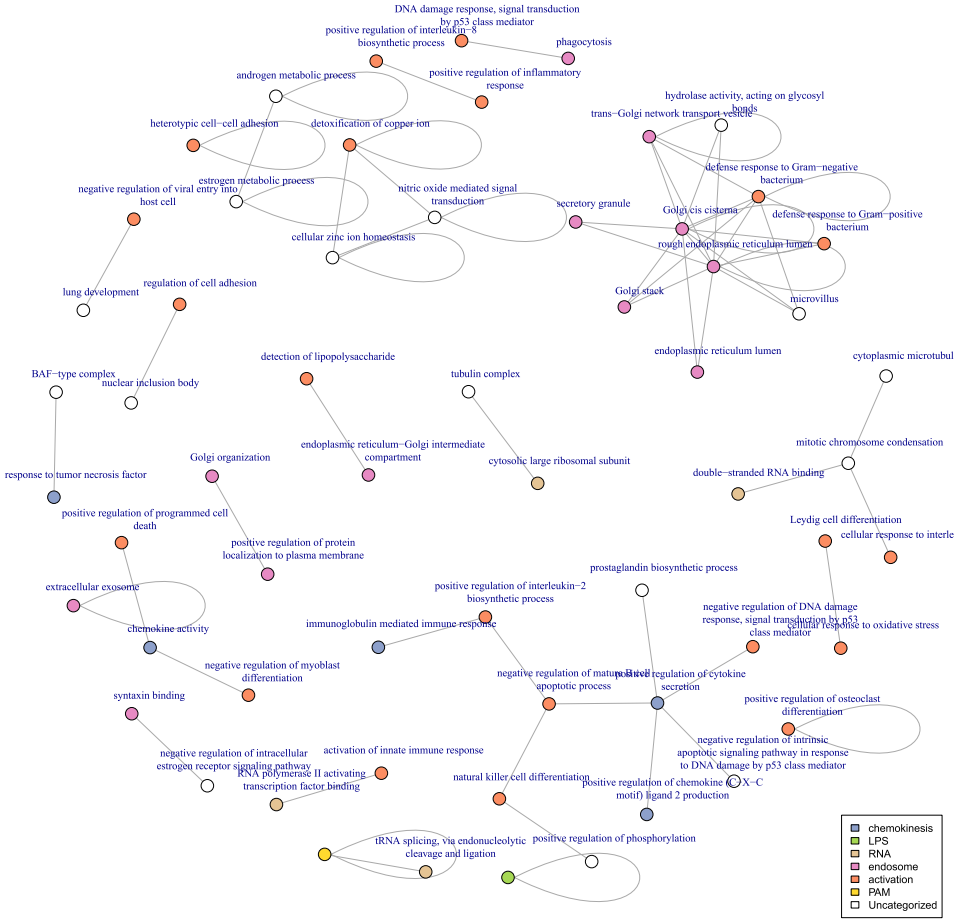

FIG. 11. *Modules enriched at FDR* $\leq$ 10% *using graphical geneset edge enrichment in mouse dendritic cells under Hurdle model.*

connections, including defense response to Gram-negative bacteria, and cell–cell adhesion and several modules involving extracellular secretion via the Golgi apparatus. Also of particular note, genes annotated to the activation of innate immune response are directly connected to RNA PolII transcription factors, as well as "detection of lipopolysaccharide"—"endoplasmic reticulum—Golgi intermediate compartment." Both of the modules are absent from the Gaussian network. This suggests that the more appropriate Hurdle model manages to identify transcription factor-induced expression changes in these regulated genes, a direct method by which one gene would induce expression changes in another.

No significant enrichment was found in the logistic model.

8. Discussion. Graphical models estimated from single cell data are distinct from networks estimated from bulk data, or even repeated stochastic samples. In

simulations, the Hurdle model with anisometric penalty has much greater sensitivity compared to available methods, while in the two data sets here, it yields substantially different network estimates compared to Gaussian and Logistic models on these zero-inflated data. When enrichment of gene ontology categories is considered between vertices in transcriptome-wide data, the enrichment uncovered with the Hurdle model is consistent with identifying direct effects of transcription factors on genes undergoing dynamic regulation due to LPS exposure.

In our work, we have utilized methods for sparse neighborhood selection. However, the zero-inflated parametric model explored here is not limited to this framework, and could serve as a basis for many network inference techniques, including mutual information-based techniques or to parametrize families of directed networks.

Although measuring transcriptome-wide data allows conditional estimation of direct effects between genes, non-mRNA factors may also greatly affect gene expression. In this sense, important variables have still been marginalized over, and in the case of the Tfh data, indeed, most of the transcriptome has been marginalized over. Extensions that adapt graphical model selection to clustering and/or factor analytic models would likely be useful and allow greater biological insight with these data sets.

Acknowledgments. AM thanks Daniel Lu for comments on the networks in Section 6.

SUPPLEMENTARY MATERIAL

Derivations and methods (DOI: 10.1214/18-AOAS1213SUPP; .zip). Supplemental derivations and methods for simulation and data preprocessing.

REFERENCES

ADACHI, Y., HIRAMATSU, S., TOKUDA, N., SHARIFI, K., EBRAHIMI, M., ISLAM, A., KAGAWA, Y., KOSHY VAIDYAN, L., SAWADA, T., HAMANO, K. and OWADA, Y. (2012). Fatty acid-binding protein 4 (FABP4) and FABP5 modulate cytokine production in the mouse thymic epithelial cells. *Histochem. Cell Biol.* **138** 397–406.

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. MR3335095

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193

THE GENE ONTOLOGY CONSORTIUM Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43**. (D1): D1049–D1056, 2015.

DE JONG, E. C., VIEIRA, P. L., KALINSKI, P., SCHUITEMAKER, J. H. N., TANAKA, Y., WIERENGA, E. A., YAZDANBAKHSH, M. and KAPSENBERG, M. L. (2002). Microbial compounds selectively induce Th1 cell-promoting or Th2 cell-promoting dendritic cells in vitro with diverse th cell-polarizing signals. *J. Immunol.* **168** 1704–1709.

DENDA-NAGAI, K., AIDA, S., SABA, K., SUZUKI, K., MORIYAMA, S., OO-PUTHINAN, S., TSUIJI, M., MORIKAWA, A., KUMAMOTO, Y., SUGIURA, D., KUDO, A., AKIMOTO, Y., KAWAKAMI, H., BOVIN, N. V. and IRIMURA, T. (2010). Distribution and function of macrophage galactose-type C-type lectin 2 (MGL2/CD301b): Efficient uptake and presentation of glycosylated antigens by dendritic cells. *J. Biol. Chem.* **285** 19193–19204.

DOBRA, A., HANS, C., JONES, B., NEVINS, J. R., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941

DRTON, M. and MAATHUIS, M. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393.

DRTON, M., STURMFELS, B. and SULLIVANT, S. (2009). *Lectures on Algebraic Statistics*. *Oberwolfach Seminars* **39**. Birkhäuser, Basel. MR2723140

ELTOFT, T., KIM, T. and LEE, T. W. (2006). On the multivariate Laplace distribution. *IEEE Signal Process. Lett.* **13** 300–303.

FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., JULIANA MCELRATH, M., PRLIC, M., LINSLEY, P. S. and GOTTARDO, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.

FOYGEL, R. and DRTON, M. (2010). Exact block-wise optimization in group lasso and sparse group lasso for linear regression. 1–19. Arxiv preprint. Available at arXiv:1010.3320.

MARINOV, G. K., WILLIAMS, B. A., MCCUE, K., SCHROTH, G. P., GERTZ, J., MYERS, R. M. and WOLD, B. J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24** 496–510.

HERMANN-KLEITER, N. and BAIER, G. (2010). NFAT pulls the strings during CD4+ T helper cell effector functions. Unpublished manuscript.

JANES, K. A., WANG, C.-C., HOLMBERG, K. J., CABRAL, K. and BRUGGE, J. S. (2010). Identifying single-cell molecular programs by stochastic profiling. *Nat. Methods* **7** 311–317.

JOHNSTON, R. J., POHOLEK, A. C., DITORO, D., YUSUF, I., ETO, D., BARNETT, B., DENT, A. L., CRAFT, J. and CROTTY, S. (2009). Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. *Science* **325**.

KIM, J. K. and MARIONI, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**.

PHAM, L. V., TAMAYO, A. T., YOSHIMURA, L. C., LIN-LEE, Y. C. and FORD, R. J. (2005). Constitutive NF-kappaB and NFAT activation in aggressive B-cell lymphomas synergistically activates the CD154 gene and maintains lymphoma cell survival. *Blood* **106** 3940–3947.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford University Press, New York. MR1419991

LEE, J. D. and HASTIE, T. J. (2013). Structure learning of mixed graphical models. In *AISTATS* 16 **31** 388–396, Scottsdale, AZ. Available at http://jmlr.org/proceedings/papers/v31/lee13a.html.

LI, Y., PEARL, S. A. and JACKSON, S. A. (2015). Gene networks in plant biology: Approaches in reconstruction and analysis. *Trends Plant Sci.* **20** 664–675.

LIN, L., FINAK, G., USHEY, K., SESHADRI, C., HAWN, T. R., FRAHM, N., SCRIBA, T. J., MAHOMED, H., HANEKOM, W. et al. (2015). COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nat. Biotechnol.* **33** 610–616.

MA, C. S., DEENICK, E. K., BATTEN, M. and TANGYE, S. G. (2012). The origins, function, and regulation of T follicular helper cells. *J. Exp. Med.* **209** 1241–1253.

MARKOWETZ, F. and SPANG, R. (2007). Inferring cellular networks: A review. *BMC Bioinform.* **8**.

MCDAVID, A., FINAK, G., CHATTOPADYAY, P. K., DOMINGUEZ, M., LAMOREAUX, L., MA, S. S., ROEDERER, M. and GOTTARDO, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29** 461–467.

McDavid, A., Gottardo, R., Simon, N. and Drton, M. (2019). Supplement to "Graphical models for zero-inflated single cell gene expression." DOI:10.1214/18-AOAS1213SUPP.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.

Precopio, M. L., Betts, M. R., Parrino, J., Price, D. A., Gostick, E., Ambrozak, D. R., Asher, T. E., Douek, D. C., Harari, A. et al. (2007). Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. *J. Exp. Med.* **204** 1405–1416.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T. et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510** 263–269.

Simon, N. and Tibshirani, R. (2012). Standardization and the group Lasso penalty. *Statist. Sinica* **22** 983–1001.

Tansey, W., Padilla, O. H. M., Suggala, A. S. and Ravikumar, P. (2015). Vector-space Markov random fields via exponential families. In *Proceedings of the* 32*nd International Conference on Machine Learning* **37** 684–692. Available at http://jmlr.org/proceedings/papers/v37/tansey15.html.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 245–266. MR2899862

Yang, E., Baker, Y., Ravikumar, P., Allen, G. and Liu, Z. (2014). Mixed graphical models via exponential families. In *AISTATS* 17 **33**. Reykjavik, Iceland.

A. McDavid
Department of Biostatistics
and Computational Biology
University of Rochester Medical Center
265 Crittenden Blvd
Rochester, New York 14642
USA
E-mail: andrew_mcdavid@urmc.rochester.edu

R. Gottardo
Vaccine and Infectuous Disease Division
Fred Hutchinson Cancer Research Center
and
Department of Statistics
University of Washington
Box 354322
Seattle, Washington 98195-4322
USA
E-mail: rgottard@fredhutch.org

N. Simon
Department of Biostatistics
University of Washington
Box 357232
Seattle, Washington 98195-7232
USA
E-mail: nrsimon@uw.edu

M. Drton
Department of Statistics
University of Washington
Box 357232
Seattle, Washington 98195-4322
USA
and
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5, 2100 København
Denmark
E-mail: md5@uw.edu