Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

A Markov Random Field Model for Network-based Differential Expression Analysis of Single-cell RNA-seq Data

Hongyu Li [1], Zhichao Xu [1], Taylor Adams [2], Naftali Kaminski [2] and Hongyu Zhao [1,*]

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT 06511, USA and
[2] Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT 06520, USA

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Recent development of single cell sequencing technologies has made it possible to identify genes with different expression (DE) levels at the cell type level between different groups of samples. However, the often-low sample size of single cell data limits the statistical power to identify DE genes.
Results: In this article, we propose to borrow information through known biological networks. Our approach is based on a Markov Random Field (MRF) model to appropriately accommodate gene network information as well as dependencies among cells to identify cell-type specific DE genes. We implement an Expectation-Maximization (EM) algorithm with mean field-like approximation to estimate model parameters and a Gibbs sampler to infer DE status. Simulation study shows that our method has better power to detect cell-type specific DE genes than conventional methods while appropriately controlling type I error rate. The usefulness of our method is demonstrated through its application to study the pathogenesis and biological processes of idiopathic pulmonary fibrosis (IPF) using a single-cell RNA-sequencing (scRNA-seq) data set, which contains 18,150 protein-coding genes across 38 cell types on lung tissues from 32 IPF patients and 28 normal controls.
Availability: The algorithm is implemented in R. The source code can be downloaded at https://github.com/eddiehli/MRFscRNAseq.
Contact: hongyu.zhao@yale.edu
Supplementary information: Supplementary data are available online.

1 Introduction

With recent advancement in single-cell RNA sequencing (scRNA-seq) technologies, it has opened up unique opportunities to understand genomic and proteomic changes at the single cell resolution. Such data allow us to identify cell type specific DE genes that are associated with diseases, e.g. idiopathic pulmonary fibrosis (IPF). Over the last decade, many statistical methods have been proposed to detect DE genes using scRNA-seq data. For example, Single-Cell Differential Expression (SCDE) (Kharchenko *et al.*, 2014) fits a mixture of Poisson model and Negative Binominal model to model the zeros and positive mean expressions separately. Model-based Analysis of Single-cell Transcriptomics (MAST) (Finak *et al.*, 2015) utilizes a two-part hurdle model to simultaneously model the rate of expression and the mean expression level. One novel aspect of MAST is that it adjusts the fraction of genes expressed across cells to obtain more reliable estimates. scDD (Korthauer *et al.*, 2016) uses a conjugate Dirichlet process mixture to identify DE genes. Meanwhile, nonparametric methods such as EMDomics (Nabavi *et al.*, 2016) and D3E (Delmans and Hemberg, 2016) have also been developed for scRNA-seq data to detect DE genes. Nonetheless, although it is well known that genes and cells do not work independently, none of the existing methods take gene network

information or dependencies among cells into consideration. Given the large scale and complexity of the scRNA-seq data, one key challenge is how to appropriately accommodate these dependencies to better identify cell-type specific DE genes.

In this paper, we propose a Markov Random Field (MRF) model that can capture gene network and cell type information. We note that the MRF model has been applied to both genome-wide association studies and bulk RNA-seq studies to model dependencies in genomic and transcriptomic data. For instance, biological pathways were used to model the structure of neighboring genes (Wei and Li, 2007; Wei and Li, 2008; Chen *et al.*, 2011). In addition, the similarities between brain regions and adjacent time points were incorporated to jointly model the spatial-temporal dependencies for human neurodevelopment data (Lin *et al.*, 2015). Our method adopts local false discovery rate framework that was developed by Efron (Efron, 2004) to identify cell-type specific DE genes. We implement an efficient EM algorithm with mean field-like approximation (Zhang, 1992; Celeux *et al.*, 2003; Lin *et al.*, 2016) to estimate model parameters and Gibbs sampler to estimate the posterior probabilities. In Section 2, we introduce the MRF model, parameter estimation, and posterior probability estimation. Section 3 shows results from simulation studies. We apply our method to analyze the lung IPF scRNA-seq data to identify DE genes in Section 4. We conclude the manuscript with a brief discussion in Section 5.

2 Methods

2.1 Model Setup

Given normalized single cell RNA-seq data, y_{gcpk} denotes the normalized observed expression of gene g in cell type c in the kth replicate in condition p. For simplicity, we assume $P = 2$. We assume that the cells have been correctly assigned to their corresponding cell types. In each group, there are n_{gcp} samples for the pth group (either disease or control group). Let $\mathbf{y}_{gc1}$ and $\mathbf{y}_{gc2}$ denote the vectors of expression values for gene g in cell type c for the two groups. The two-sample t-statistic is then

$$t_{gc} = \frac{\bar{\mathbf{y}}_{gc1} - \bar{\mathbf{y}}_{gc2}}{\text{se}\left(\bar{\mathbf{y}}_{gc1} - \bar{\mathbf{y}}_{gc2}\right)}$$

Then we transform the test statistic into z-scores,

$$z_{gc} = \Phi^{-1}\left(F_{n_{gc1}+n_{gc2}-2}\left(t_{gc}\right)\right)$$

where n_{gc1} and n_{gc2} are the number of samples for the two groups, e.g. disease and control groups, for gene g in cell type c; Φ is the cumulative distribution function of a standard normal distribution; and F is the cumulative distribution function for a student-t distribution with $n_{gc1} + n_{gc2} - 2$ degrees of freedom. The gene expression data are then represented by a summary statistic matrix $\mathbf{Z}$, where each entry z_{gc} represents the evidence of differential expression between the two groups for each gene across cell types. Let w_{gc} denote the binary latent state representing whether gene g is deferentially expressed in cell type c between the two groups. Then $\mathbf{W}$ is the latent state matrix. Because w_{gc} has two states, we assume that z_{gc} follows a mixture distribution,

$$f\left(z_{gc} \,|\, w_{gc}\right) = \left(1 - w_{gc}\right) f_0\left(z_{gc}\right) + w_{gc}\, f_1\left(z_{gc}\right)$$

where $f_0\left(z_{gc}\right)$ is the null density and $f_1\left(z_{gc}\right)$ is the non-null density. We further assume that the null density follows a standard normal distribution. The non-null density is estimated through Efron's nonparametric empirical Bayes framework (Efron, 2004). The inference on the latent state $\mathbf{W}$ is our primary objective. In the following, we construct the MRF model that accommodates cell type dependencies and gene network information. A gene network information is represented by an undirected graph, with a set of nodes $\mathcal{V}_g$, which correspond to cell-type specific genes, and a set of edges $\mathcal{E}_g$, which represent the relationships among the nodes. For each gene g, we can use the following vector to denote its cell-type specific DE status

$$\mathcal{V}_g = \{W_{gc} : c = 1, \cdots, C\}$$

The set of edges $\mathcal{E}_g$ can be divided into two subsets, $\mathcal{E}_{g1}$ and $\mathcal{E}_{g2}$. For two genes g and g', if there is a known relationship, e.g. from a pathway database, we write $g \sim g'$. For a given gene g, let $N_g = \{g' : g \sim g' \in \mathcal{E}_{g1}\}$ be the set of genes that have known relationships with gene g. Similarly, for two cell types c and c', if there is a known relationship, we write $c \sim c'$. For a given cell type c, let $N_c = \{c' : c \sim c' \in \mathcal{E}_{g2}\}$ be the set of cell types that have close relationships with cell type c. Then we can write two sets of edges as

$$\mathcal{E}_{g1} = \left\{\left(w_{gc}, w_{g'c'}\right) : g \sim g', c = c'\right\}$$
$$\mathcal{E}_{g2} = \left\{\left(w_{gc}, w_{g'c'}\right) : g = g', c \sim c'\right\}$$

Therefore, edges in $\mathcal{E}_{g1}$ capture similarities between genes based on gene network information, while edges in $\mathcal{E}_{g2}$ capture the dependencies between cell types. Then we construct a pairwise interaction MRF model (Besag, 1986),

$$\begin{aligned} p\left(\mathbf{W} \,|\, \mathbf{\Phi}\right) \propto \prod_{g=1}^{G} \exp\Bigg\{ & \gamma_0 \sum_{\mathcal{V}_g} \mathcal{I}_0\left(w_{gc}\right) + \gamma_1 \sum_{\mathcal{V}_g} \mathcal{I}_1\left(w_{gc}\right) + \\ & \beta_{\text{gene}} \sum_{\mathcal{E}_{g1}} \Big[\mathcal{I}_0\left(w_{gc}\right)\mathcal{I}_0\left(w_{g'c'}\right) + \mathcal{I}_1\left(w_{gc}\right)\mathcal{I}_1\left(w_{g'c'}\right)\Big] + \\ & \beta_{\text{cell}} \sum_{\mathcal{E}_{g2}} \Big[\mathcal{I}_0\left(w_{gc}\right)\mathcal{I}_0\left(w_{g'c'}\right) + \mathcal{I}_1\left(w_{gc}\right)\mathcal{I}_1\left(w_{g'c'}\right)\Big]\Bigg\} \end{aligned}$$

Let $\gamma = \gamma_1 - \gamma_0$, the conditional probability for the cell-type specific DE status is

$$p\left(w_{gc} \,|\, \mathbf{W} \,/\, w_{gc}; \mathbf{\Phi}\right) = \frac{\exp\left\{w_{gc}\, F\left(w_{gc}, \mathbf{\Phi}\right)\right\}}{1 + \exp\left\{F\left(w_{gc}, \mathbf{\Phi}\right)\right\}} \tag{1}$$

where

$$F\left(w_{gc}, \mathbf{\Phi}\right) = \gamma + \beta_{\text{gene}} \sum_{g' \in N_g} \left(2w_{g'c} - 1\right) + \beta_{\text{cell}} \sum_{c' \in N_c} \left(2w_{gc'} - 1\right)$$

where "/" denotes other than; $\mathbf{\Phi} = \left(\gamma, \beta_{\text{gene}}, \beta_{\text{cell}}\right)$. Here β_{gene} is the parameter that captures the similarities between genes, and β_{cell} is the parameter that captures cell type dependencies.

2.2 Parameter Estimation

For parameter estimation, we utilize the EM algorithm with mean field-like approximation (Zhang, 1992; Celeux *et al.*, 2003; Lin *et al.*, 2016). Let $\tilde{\mathbf{W}}$ denote a configuration, the joint distribution $p\left(\mathbf{W} \,|\, \mathbf{\Phi}\right)$ can be estimated by

$$p_{\tilde{\mathbf{W}}}\left(\mathbf{W} \,|\, \mathbf{\Phi}\right) = \prod_{c=1}^{C} \prod_{g=1}^{G} p\left(w_{gc} \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right); \mathbf{\Phi}\right)$$

where $\tilde{\mathcal{N}}\left(w_{gc}\right)$ represents the neighbors of w_{gc} corresponding to the fixed configuration $\tilde{\mathbf{W}}$. The complete log likelihood is

$$\log p_{\tilde{\mathbf{W}}}\left(\mathbf{W}, \mathbf{Z} \,|\, \mathbf{\Phi}\right) = \sum_{g,c} \log p\left(z_{gc} \,|\, w_{gc}\right) + \sum_{g,c} \log p\left(w_{gc} \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right); \mathbf{\Phi}\right)$$

The Q function is

Table 1. Comparison of Performance for t-test, Hurdle Model and the Proposed MRF Model on Simulated Data

Number of DE Pathways	Gamma Parameters	τ	Model	Sensitivity	Specificity	FDR
$K = 5$ DE Proportion: 13%	$\alpha = 2$ $\beta = 4$	$\tau = 5$	t-test	0.4776 (0.0017)	0.9992 (0.0000)	0.0223 (0.0005)
			Hurdle	0.4935 (0.0016)	0.9969 (0.0000)	0.0827 (0.0011)
			MRF	0.5069 (0.0105)	0.9992 (0.0001)	0.0215 (0.0012)
		$\tau = 10$	t-test	0.5709 (0.0016)	0.9991 (0.0000)	0.0231 (0.0006)
			Hurdle	0.5158 (0.0014)	0.9968 (0.0000)	0.0827 (0.0011)
			MRF	0.7482 (0.0102)	0.9977 (0.0001)	0.0411 (0.0016)
	$\alpha = 3$ $\beta = 5$	$\tau = 5$	t-test	0.5044 (0.0017)	0.9991 (0.0000)	0.0237 (0.0005)
			Hurdle	0.5320 (0.0015)	0.9966 (0.0000)	0.0830 (0.0010)
			MRF	0.5820 (0.0120)	0.9990 (0.0001)	0.0234 (0.0011)
		$\tau = 10$	t-test	0.5915 (0.0014)	0.9990 (0.0000)	0.0241 (0.0006)
			Hurdle	0.5344 (0.0016)	0.9966 (0.0000)	0.0837 (0.0010)
			MRF	0.7726 (0.0117)	0.9975 (0.0001)	0.0425 (0.0013)
$K = 10$ DE Proportion: 24%	$\alpha = 2$ $\beta = 4$	$\tau = 5$	t-test	0.5134 (0.0012)	0.9985 (0.0000)	0.0217 (0.0004)
			Hurdle	0.5201 (0.0011)	0.9949 (0.0001)	0.0680 (0.0008)
			MRF	0.6244 (0.0160)	0.9992 (0.0000)	0.0104 (0.0006)
		$\tau = 10$	t-test	0.6171 (0.0010)	0.9981 (0.0000)	0.0227 (0.0004)
			Hurdle	0.5376 (0.0010)	0.9948 (0.0001)	0.0675 (0.0007)
			MRF	0.8382 (0.0126)	0.9971 (0.0001)	0.0243 (0.0008)
	$\alpha = 3$ $\beta = 5$	$\tau = 5$	t-test	0.5417 (0.0012)	0.9984 (0.0000)	0.0221 (0.0005)
			Hurdle	0.5628 (0.0012)	0.9943 (0.0001)	0.0703 (0.0007)
			MRF	0.7206 (0.0143)	0.9992 (0.0000)	0.0085 (0.0005)
		$\tau = 10$	t-test	0.6416 (0.0012)	0.9980 (0.0000)	0.0234 (0.0004)
			Hurdle	0.5633 (0.0012)	0.9944 (0.0001)	0.0694 (0.0007)
			MRF	0.8658 (0.0104)	0.9962 (0.0001)	0.0308 (0.0008)
$K = 18$ DE Proportion: 34%	$\alpha = 2$ $\beta = 4$	$\tau = 5$	t-test	0.5365 (0.0010)	0.9978 (0.0000)	0.0195 (0.0004)
			Hurdle	0.5367 (0.0010)	0.9933 (0.0001)	0.0574 (0.0006)
			MRF	0.6279 (0.0172)	0.9992 (0.0000)	0.0072 (0.0005)
		$\tau = 10$	t-test	0.6457 (0.0010)	0.9971 (0.0000)	0.0214 (0.0003)
			Hurdle	0.5519 (0.0009)	0.9931 (0.0001)	0.0572 (0.0006)
			MRF	0.8690 (0.0102)	0.9962 (0.0001)	0.0206 (0.0007)
	$\alpha = 3$ $\beta = 5$	$\tau = 5$	t-test	0.5646 (0.0010)	0.9976 (0.0000)	0.0202 (0.0003)
			Hurdle	0.5821 (0.0011)	0.9923 (0.0001)	0.0607 (0.0005)
			MRF	0.7503 (0.0145)	0.9994 (0.0000)	0.0044 (0.0003)
		$\tau = 10$	t-test	0.6744 (0.0010)	0.9970 (0.0000)	0.0216 (0.0003)
			Hurdle	0.5869 (0.0011)	0.9922 (0.0001)	0.0606 (0.0005)
			MRF	0.8991 (0.0046)	0.9946 (0.0001)	0.0283 (0.0006)

* Summaries are averaged over 100 simulations; standard errors are shown in parentheses.

$$Q\left(\mathbf{\Phi} \,|\, \hat{\mathbf{\Phi}}\right) = \sum_{g,c} \Big\{ p\left(w_{gc} = 0 \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right), \mathbf{Z}; \hat{\mathbf{\Phi}}\right) \log p\left(w_{gc} = 0 \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right); \hat{\mathbf{\Phi}}\right) + p\left(w_{gc} = 1 \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right), \mathbf{Z}; \hat{\mathbf{\Phi}}\right) \log p\left(w_{gc} = 1 \,|\, \tilde{\mathcal{N}}\left(w_{gc}\right); \hat{\mathbf{\Phi}}\right) \Big\}$$

Then we use the following EM algorithm to estimate the model parameters

1. Estimate f_0 and f_1 using R package `locfdr` based on the z-scores. Then obtain $\tilde{\mathbf{W}}$ using the mixture model;
2. Expectation-step: Let $\hat{\mathbf{\Phi}}^{(k)}$ be the estimated parameters in the kth iteration. The Q function $Q(\mathbf{\Phi} \,|\, \hat{\mathbf{\Phi}}^{(r)})$ can be calculated from the fixed configuration $\tilde{\mathbf{W}}$;
3. Maximization-step: Update $\mathbf{\Phi}$ with $\hat{\mathbf{\Phi}}^{(k+1)}$, which maximizes $Q(\mathbf{\Phi} \,|\, \hat{\mathbf{\Phi}}^{(k)})$;
4. Update $\tilde{w}_{gc}$ by the Gibbs sampler with posterior probability proportional to $p(z_{gc} \,|\, \tilde{w}_{gc})\, p(\tilde{w}_{gc} \,|\, \mathcal{N}(\tilde{w}_{gc}); \hat{\mathbf{\Phi}}^{(k+1)})$, and obtain $\tilde{\mathbf{W}}$;
5. Repeat steps 2 – 4 until convergence.

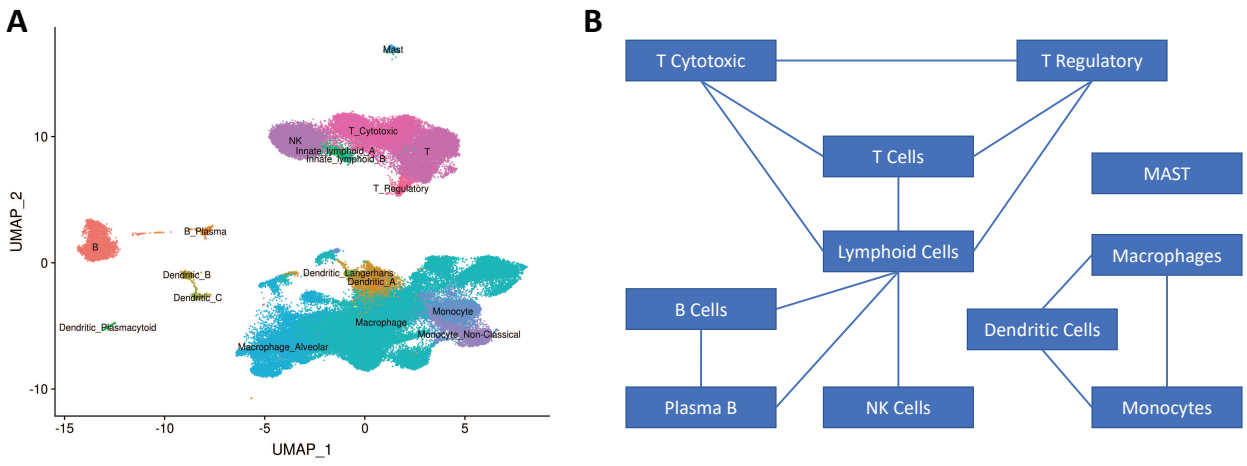

Fig. 1. Cell Type Clusters. (A) shows the UMAP of eighteen immune cell types, and (B) shows the network among these immune cell types that are determined by domain knowledge.

2.3 Detecting Differentially Expressed Genes

We use the posterior probability-based definition of FDR (Newton *et al.*, 2001; Li *et al.*, 2010) to identify DE genes. We first estimate the posterior local FDR p_{gc} using Gibbs sampler,

$$p_{gc} = p\left(w_{gc} = 0 \,|\, \mathbf{Z}; \hat{\mathbf{\Phi}}\right)$$

Then we sort p_{gc} in ascending order, and let $p_{(i)}$ denote the sorted values. We find k such that

$$k = \max\left\{ j : \frac{1}{j}\sum_{i=1}^{j} p_{(i)} \leq \alpha \right\}$$

and we reject the first k null hypotheses.

3 Simulation Study

Simulation study was conducted to assess the performance of our proposed MRF procedure. We obtained the network structure of 52 pathways (Supplementary Table 1) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). We only considered gene-gene relations, and excluded any gene-compound relations and compound-compound relations. In total, there were 965 genes (nodes) and 5155 interactions (edges) in these 52 pathways. The number of cell types was set to be 20. We randomly selected K pathways, and set all genes in these K pathways in the first 10 cell types to be DE. The rest were set to be equally expressed. The latent states were simulated by a Gibbs sampler with 10 rounds. In each iteration, the latent state W was sequentially updated according to Equation (1). Then the counts were simulated from a Negative Binomial distribution with mean μ and dispersion $\phi = 0.1$. Two groups were considered. For the control group, the mean expression levels were generated from a Gamma distribution, $\mu_{1gc} \sim \text{Gamma}(\alpha, \beta)$. For the case group, if $w_{gc} = 0$, then $\mu_{2gc} = \mu_{1gc}$; otherwise, $\mu_{2gc} = \lambda \cdot \mu_{1gc}$, where $\lambda = \tau$ or $1/\tau$ with equal probability. Four combinations of α, β and τ were considered,

- $\mu_{1gc} \sim \text{Gamma}\,(2, 4)$ and $\tau = 5$
- $\mu_{1gc} \sim \text{Gamma}\,(2, 4)$ and $\tau = 10$
- $\mu_{1gc} \sim \text{Gamma}\,(3, 5)$ and $\tau = 5$
- $\mu_{1gc} \sim \text{Gamma}\,(3, 5)$ and $\tau = 10$

We set $K = 5, 10, 18$ to reflect different proportions of DE genes in each setting. We compared our results with two-sample t-tests and a hurdle model that simultaneously modeled the rate of expression and the mean expression level. The simulation results are shown in Table 1. These three methods had comparable specificity and false discovery rates. However, our proposed MRF model performed significantly better than the other two methods in terms of sensitivity, especially in the settings when $\tau = 10$.

4 Data Application

We applied our method to a recent study that collected single-cell RNA-sequencing (scRNA-seq) data using lung tissues from 32 IPF patients and 28 normal controls (Adams *et al.*, 2019). A total of 18,150 protein-coding genes were profiled across 38 cell types.

Idiopathic pulmonary fibrosis (IPF) is an incurable aggressive lung disease. It progressively scars the lung and causes usual interstitial pneumonia (UIP). However, to date, what causes the scarring remains unknown. IPF affects around three million people globally (Martinez *et al.*, 2017, Barratt *et al.*, 2018), with its mortality rate much higher than many cancers, and the median survival time for patients without a lung transplant is about three to four years (Ley *et al.*, 2011, Raghu *et al.*, 2014). Many efforts have been made to understand the pathogenesis and biological processes of this disease. For instance, genome-wide association studies (GWAS) have identified 20 regions in the human genome that are associated with increased risk to IPF (Fingerlin *et al.*, 2013, Noth *et al.*, 2013, Allen *et al.*, 2017, Allen *et al.*, 2020). In addition, transcriptome analyses through microarrays (Zuo *et al.*, 2002, Meltzer *et al.*, 2011, Yang *et al.*, 2013, Yue *et al.*, 2013) and RNA-seq (Deng *et al.*, 2013, Nance *et al.*, 2014, McDonough *et al.*, 2019) have revealed several genes and pathways that are related to IPF. In particular, a recent review article described in details how transcriptome analyses helped to identify novel genes involved in the pathogenesis of IPF and the importance of using single-cell RNA-seq analysis to discover cell-type specific DE genes (Vukmirovic and Kaminski, 2018).

We used Seurat (Butler *et al.*, 2018) to perform data pre-processing and quality control. Specifically, cells that had unique gene counts greater than 5,000 or less than 200 were filtered out. Cells that had more than 5% mitochondrial counts were also excluded from further analysis. In total, there were 18,150 protein-coding genes across 114, 364 cells remaining after quality control. We normalized the expression data for each cell by the total expression multiplied by a scale factor of 10,000 and then log-transformed the results. Since about 87% of the cells were myeloid and

lymphoid cells, we focused on the immune cells in further analyses. There were 18 distinct immune cell types (Figure 1A).

In this paper, instead of considering a network with 18,150 genes across 18 cell types, we focused on 2,000 genes that exhibited high cell-to-cell variation between cell types. Previous research (Brennecke *et al.*, 2013, Butler *et al.*, 2018) showed that focusing on these highly variable genes in DE analysis helps to highlight significant biological signals. We extracted the gene network information from two well-known protein-protein interaction network (PPIN) databases, BioGrid (Oughtred *et al.*, 2019) and IntAct (Orchard *et al.*, 2014). For these 2,000 highly variable genes, the BioGrid database identified 5,400 edges, while 3,104 edges were found in the IntAct database. The two gene networks were visualized in Supplementary Figures 1 and 2. There was an overlap of 1,754 edges between the two databases. In addition, the dependencies among cells were determined by domain knowledge (Figure 1B).

Table 2. Comparison of Performance for t-test, Hurdle Model and the Proposed MRF Model on Simulated Data

Input Statistics	**Models**	**# DEGs**[a]	**# Cell-type Specific DEGs**[b]
t-test	w/o MRF	1472	3607
	MRF w/ BioGrid	1602	5042
	MRF w/ IntAct	1597	5033
Wilcoxon	w/o MRF	1562	3945
	MRF w/ BioGrid	1752	6525
	MRF w/ IntAct	1757	6494
MAST	w/o MRF	1721	4826
	MRF w/ BioGrid	1863	8731
	MRF w/ IntAct	1864	8707
Bimodal	w/o MRF	1740	5159
	MRF w/ BioGrid	1889	9648
	MRF w/ IntAct	1890	9431

[a] Number of genes that were found DE in at least one cell type;
[b] Aggregated number of cell-type specific DEGs.

For these 2,000 genes across 18 cell types, we fitted two separate MRF models utilizing gene networks from BioGrid and IntAct, respectively. First, we obtained the parameter estimates by implementing the EM algorithm with 200 iterations. With parameter estimates fixed, we then ran 20,000 iterations of Gibbs sampler with 10,000 iterations as burn-in to obtain posterior probabilities. For MRF parameters, β_{gene} was estimated to be 0.05 with BioGrid gene network, whereas β_{gene} was estimated to be around 0.08 with IntAct gene network. For β_{cell}, both models yielded estimated values around 0.19. We note that β_{cell} was much larger than β_{gene}, which suggested stronger cell-type dependency than that through gene network information. For DE analysis, the threshold for posterior probability was set at 0.92 for both models with $\alpha = 0.01$. Out of 2,000 genes across 18 cell types, the MRF model with BioGrid gene network identified 1,602 genes that were found DE in at least one cell type. For the IntAct gene network, the MRF model identified 1,597 genes that were DE in at least one cell type. We compared these results with two-sample t-tests using the Benjamini and Hochberg's procedure for FDR, which only identified 1,472 DE genes. We also compared the number of cell-type specific DE genes inferred by the t-test and two MRF models in Figure 2 and the detailed gene lists were supplied in the Supplementary Excel file. By utilizing gene network and cell type dependency information, the MRF model successfully identified additional sets of 130 and 125 novel DE genes with BioGrid and IntAct gene network, respectively, compared with using t-test alone. In addition, there was an overlap of 120 genes between the two novel DE gene sets, which suggested that our method had fairly robust performance with different network structures.

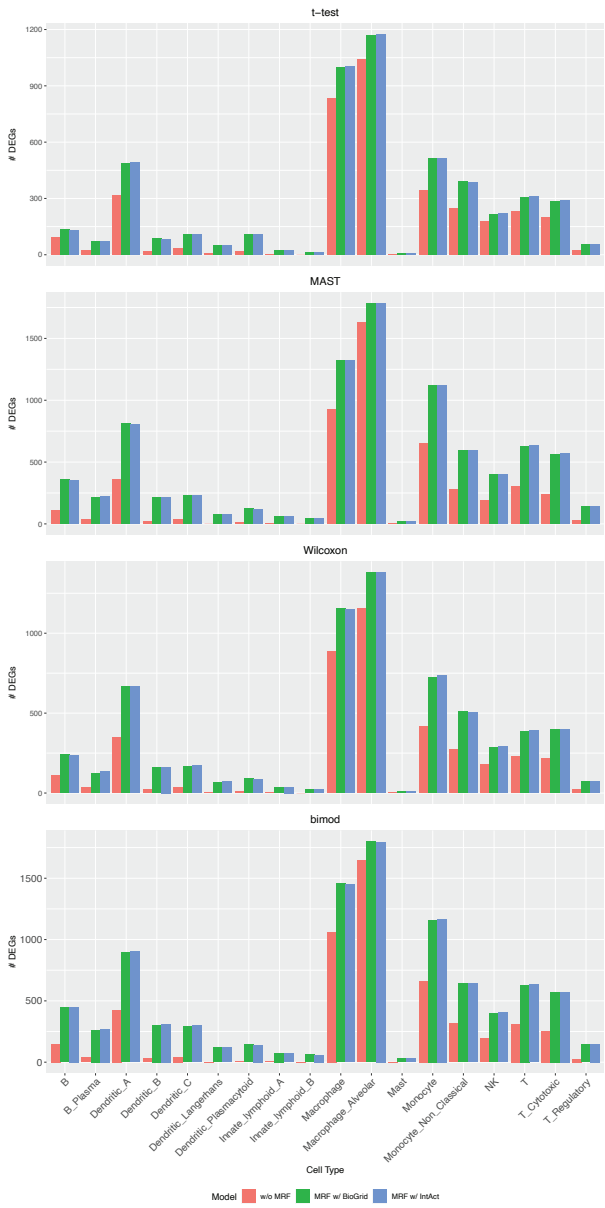

Fig. 2. Number of cell-type specific DE Genes identified using original test statistics and MRF model with BioGrid and IntAct database for four types of test statistic inputs: two-sample t-test, MAST analysis, Wilcoxon test, and likelihood ratio tests that assumes bimodal expression patterns.

Furthermore, our method can be extended to other types of test statistics in addition to the two-sample t-test. For instance, Seurat provides several built-in methods for its differential expression analysis, which typically outputs p-values and log-fold changes. Our model can be readily applied to these existing models with ease. In the IPF scRNA-seq analysis, we applied the model using the same gene sets and biological networks to the three other built-in methods: Wilcoxon test, which is the default method for

differential expression analysis in Seurat; MAST, which simultaneously models the rate of expression and the mean expression level; and a likelihood ratio test that assumes bimodal expression patterns. The results were listed in Table 2.

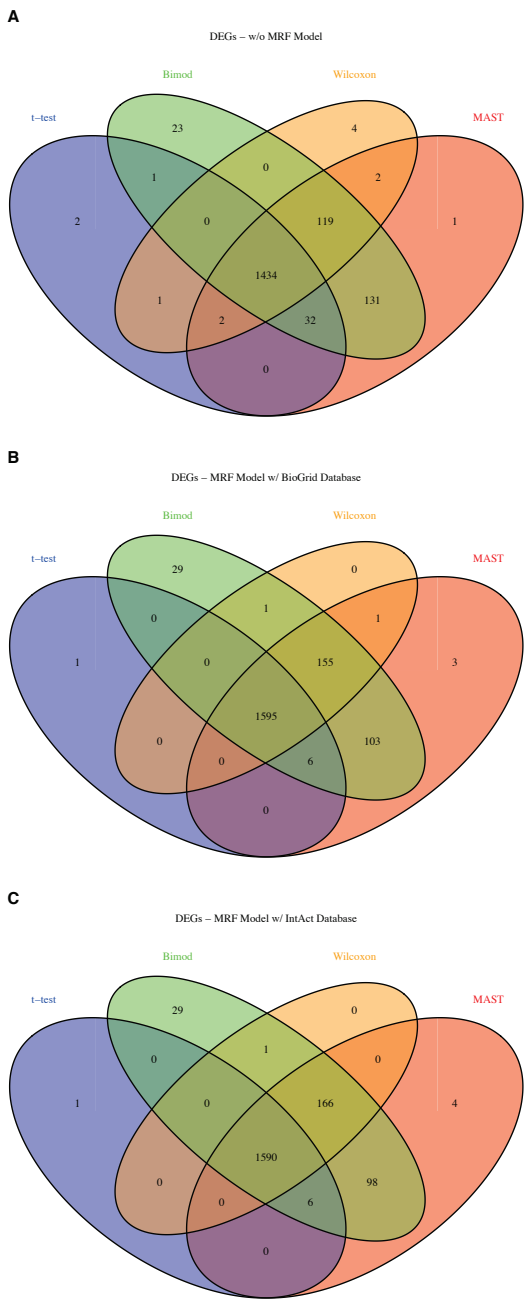

Fig. 3. Venn Diagrams of genes identified differentially expressed in at least one cell type. (A) shows the Venn Diagram of DEGs identified by four methods without the MRF model, (B) shows the Venn Diagram of DEGs identified by four methods with the MRF model utilizing the BioGrid gene network, and (C) shows the Venn Diagram of DEGs identified by four methods with the MRF model utilizing the IntAct gene network.

We compared cell-type specific DE genes inferred by the model without network information and two MRF models in Figure 2. We also compared DEGs identified across four types of test statistics. Venn diagrams of DE results inferred with or without the MRF model were visualized in Figure 3, and Venn diagrams of cell-type specific DE results were also visualized (Supplementary Figure 3). For different test statistic inputs, our model is able to identify an additional set of novel DE genes utilizing gene-gene and cell-cell networks. In addition, we note that t-test actually yields the most conservative results among the four methods.

5 Discussion

In this paper, we proposed a Markov Random Field (MRF) model to appropriately accommodate gene network information and dependencies among cell types to identify cell-type specific DE genes. In the simulation study, we constructed gene networks with KEGG pathways, while in the IPF scRNA-seq data application, we utilized two protein-protein interaction networks, BioGrid and IntAct. In fact, our method can be adapted to other networks that have similar structures as KEGG pathways, BioGrid, or IntAct PPIN. Furthermore, our model can be readily extended to many other existing DE methods with ease, such as Wilcoxon test, MAST, and others.

One caveat in our model is that the direction of changes in gene expressions is not directly incorporated in the model, which means that we are unable to differentiate whether these identified DE genes are up-regulated or down-regulated. One possible remedy is to use the sign of the original input test statistics to determine the sign of the DE results. For future work, weights could be added in our graphical model. For instance, transcription factors probably should have more weights because of their importance in gene regulation.

References

Adams, T. S. *et al.* (2019). Single Cell RNA-seq reveals ectopic and aberrant lung resident cell populations in Idiopathic Pulmonary Fibrosis. preprint, Genomics.

Allen, R. J. *et al.* (2017). Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *The Lancet Respiratory Medicine*, **5**(11), 869–880.

Allen, R. J. *et al.* (2020). Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, **201**(5), 564–574.

Barratt, S. *et al.* (2018). Idiopathic Pulmonary Fibrosis (IPF): An Overview. *Journal of Clinical Medicine*, **7**(8), 201.

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(3), 259–279.

Brennecke, P. *et al.* (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, **10**(11), 1093–1095.

Butler, A. *et al.* (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**(5), 411–420.

Celeux, G. *et al.* (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, **36**(1), 131–144.

Chen, M. *et al.* (2011). Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. *PLoS Genetics*, **7**(4), e1001353.

Delmans, M. and Hemberg, M. (2016). Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, **17**(1), 110.

Deng, N. *et al.* (2013). Detecting Splicing Variants in Idiopathic Pulmonary Fibrosis from Non-Differentially Expressed Genes. *PLoS ONE*, **8**(7), e68352.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465), 96–104.

Finak, G. *et al.* (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, **16**(1), 278.

Fingerlin, T. E. *et al.* (2013). Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nature Genetics*, **45**(6), 613–620.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. page 4.

Kharchenko, P. V. *et al.* (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, **11**(7), 740–742.

Korthauer, K. D. *et al.* (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, **17**(1), 222.

Ley, B. *et al.* (2011). Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, **183**(4), 431–440.

Li, H. *et al.* (2010). A hidden Markov random field model for genome-wide association studies. *Biostatistics*, **11**(1), 139–150.

Lin, Z. *et al.* (2015). A Markov random field-based approach to characterizing human brain development using spatial–temporal transcriptome data. *The Annals of Applied Statistics*, **9**(1), 429–451.

Lin, Z. *et al.* (2016). A Markov random field-based approach for joint estimation of differentially expressed genes in mouse transcriptome data. *Statistical Applications in Genetics and Molecular Biology*, **15**(2).

Martinez, F. J. *et al.* (2017). Idiopathic pulmonary fibrosis. *Nature Reviews Disease Primers*, **3**(1), 17074.

McDonough, J. E. *et al.* (2019). Transcriptional regulatory model of fibrosis progression in the human lung. *JCI Insight*, **4**(22), e131597.

Meltzer, E. B. *et al.* (2011). Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. *BMC Medical Genomics*, **4**(1), 70.

Nabavi, S. *et al.* (2016). EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics*, **32**(4), 533–541.

Nance, T. *et al.* (2014). Transcriptome Analysis Reveals Differential Splicing Events in IPF Lung Tissue. *PLoS ONE*, **9**(3), e92111.

Newton, M. *et al.* (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, **8**(1), 37–52.

Noth, I. *et al.* (2013). Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *The Lancet Respiratory Medicine*, **1**(4), 309–317.

Orchard, S. *et al.* (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, **42**(D1), D358–D363.

Oughtred, R. *et al.* (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, **47**(D1), D529–D541.

Raghu, G. *et al.* (2014). Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001–11. *The Lancet Respiratory Medicine*, **2**(7), 566–572.

Vukmirovic, M. and Kaminski, N. (2018). Impact of Transcriptomics on Our Understanding of Pulmonary Fibrosis. *Frontiers in Medicine*, **5**, 87.

Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**(12), 1537–1544.

Wei, Z. and Li, H. (2008). A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *The Annals of Applied Statistics*, **2**(1), 408–429.

Yang, I. V. *et al.* (2013). Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax*, **68**(12), 1114–1121.

Yue, X. *et al.* (2013). Overexpression of Sulf2 in idiopathic pulmonary fibrosis. *Glycobiology*, **23**(6), 709–719.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing*, **40**(10), 2570–2583.

Zuo, F. *et al.* (2002). Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proceedings of the National Academy of Sciences*, **99**(9), 6292–6297.