RESEARCH

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

Hongyu Li[1], Biqing Zhu[2], Zhichao Xu[1], Taylor Adams[3], Naftali Kaminski[3] and Hongyu Zhao[1,2*]

*Correspondence: hongyu.zhao@yale.edu
[1] Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06511, USA
Full list of author information is available at the end of the article

Abstract

Background: 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. In this article, we propose to borrow information through known biological networks to increase statistical power to identify differentially expressed genes (DEGs).

Results: We develop MRFscRNAseq, which is based on a Markov Random Field (MRF) model to appropriately accommodate gene network information as well as dependencies among cell types to identify cell-type specific DEGs. 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 DEGs 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.

Conclusions: The proposed MRF model is implemented in the R package MRFscRNAseq available on GitHub. By utilizing gene-gene and cell-cell networks, our method increases statistical power to detect differentially expressed genes from scRNA-seq data.

Keywords: Markov Random Field; Differential Expression; scRMA-seq

Background

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 differentially expressed genes (DEGs) that are associated with diseases, e.g. idiopathic pulmonary fibrosis (IPF). There exist many statistical methods that can perform DE analysis for scRNA-seq data. First of all, traditional statistical methods such as two-sample t-test, Wilcoxon rank sum test, logistic regression, and negative binomial regression are widely used to detect DEGs for scRNA-seq data. DE methods that are tailored for scRNA-seq data have also been developed over the last decade. For instance, Single-Cell Differential Expression (`SCDE`) [1] fits a mixture of Poisson model and Negative Binominal model for the zeros and positive

mean expressions separately. Model-based Analysis of Single-cell Transcriptomics (`MAST`) [2] 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` [3] uses a conjugate Dirichlet process mixture to identify DEGs. `DESingle` [4] adopts a zero-inflated negative binomial distribution to model count data and identify DEGs. Meanwhile, nonparametric methods such as `SigEMD` [5], `EMDomics` [6], and `D3E` [7] have also been developed for scRNA-seq data to detect DEGs. Review papers [8, 9] have pointed out that methods that were tailored for scRNA-seq data do not show significantly better performance compared to traditional methods. Surprisingly, traditional methods such as t-test and Wilcoxon test also have fairly robust performance.

In addition, several papers [10, 11, 12] have pointed out that variations between biological replicates should be properly controlled when performing DE analysis, i.e., individual effects. As a result, the family of pseudo-bulk methods are also commonly used in the DE analysis. These methods usually aggregate cell-level counts into sample-level pseudo-bulk counts, and then use methods that were originally proposed for bulk RNA-seq data to detect DEGs, such as `edgeR` [13], `DESeq2` [14], and `limma` [15]. On the other hand, methods based on mixed models [11] have also been proposed to capture the random effects for individuals. However, Crowell [10] showed in their comparison analysis that the mixed model-based methods did not perform significantly better compared to the aggregation-based pseudo-bulk methods. Moreover, these mixed model-based methods also require larger computational resources and longer computational time, which may not be worthwhile. Currently the detection of DEGs for scRNA-seq data still remains a challenge. 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 DEGs.

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 [16, 17, 18]. In addition, the similarities between brain regions and adjacent time points were incorporated to jointly model the spatial-temporal dependencies for human neurodevelopment data [19]. Our method adopts local false discovery rate framework that was developed by Efron [20] to identify cell-type specific DEGs. We implement an efficient EM algorithm [21] with mean field-like approximation [22, 23, 24] to estimate model parameters. Then we utilize Gibbs sampler to estimate the posterior probabilities to infer cell-type specific DE status.

We applied our method to a recent study that collected scRNA-seq data using lung tissues from 32 IPF patients and 28 normal controls [25]. The objective of the analysis is to detect cell-type specific DEGs between IPF patients and normal controls. 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 [26, 27], 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 [28, 29]. 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 [30, 31, 32, 33]. In addition, transcriptome analyses through microarrays [34, 35, 36, 37] and RNA-seq [38, 39, 40] have revealed genes and pathways related to IPF. In particular, a recent review described in detail 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 DEGs [41]. In order to assess the performance of our proposed MRF model, simulation study was conducted under various scenarios. The results for simulation study and the DE results for the IPF scRNA-seq analysis are shown in the third section. We conclude the manuscript with a brief discussion in the last section.

Methods

Markov Random Field Model

Model Setup

Given normalized single cell RNA-seq data, let y_{gcpk} denote the normalized observed expression of gene g in cell type c in the kth replicate in condition p. We let G denote the number of genes and C denote the number of cell types. 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). The number of samples here is the number of cells belonging to this cell type. 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 can be constructed as

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

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. $\mathbf{Z}$ is a $G \times C$ matrix. Let w_{gc} denote the binary latent state representing whether gene g is differentially expressed in cell type c between the two groups. Then $\mathbf{W}$ is the latent state matrix, which has the same dimension as

Z. Because w_{gc} has two states, we assume that z_{gc} follows a mixture distribution,

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

where $f_0\left(z_{gc}\right)$ is the null density and $f_1\left(z_{gc}\right)$ is the non-null density. The null and non-null densities are estimated through Efron's nonparametric empirical Bayes framework [20]. 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,

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

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 $N_g = \left\{g' : g \sim g' \in \mathcal{E}_{g1}\right\}$ be the set of genes that have known relationships with this gene. 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 $N_c = \left\{c' : c \sim c' \in \mathcal{E}_{g2}\right\}$ be the set of cell types that have close relationships with cell type c. Then we can write two sets of edges as

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

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 [42],

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

Here $\mathcal{I}$ is an indicator function, i.e., when $w_{gc} = 1$, $\mathcal{I}_1\left(w_{gc}\right) = 1$. Let $\gamma = \gamma_1 - \gamma_0$, the conditional probability for the cell-type specific DE status is

$$p\left(w_{gc}\,|\,\mathbf{W} \setminus w_{gc}; \mathbf{\Phi}\right) = \frac{\exp\left\{w_{gc}\, F\left(w_{gc}, \mathbf{\Phi}\right)\right\}}{1 + \exp\left\{F\left(w_{gc}, \mathbf{\Phi}\right)\right\}}, \tag{2}$$

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; $\mathbf{\Phi} = (\gamma, \beta_{\text{gene}}, \beta_{\text{cell}})$. Here β_{gene} is the parameter that captures the similarities between genes, and β_{cell} is the parameter that captures cell type dependencies.

Parameter Estimation

For parameter estimation, we adopt the EM algorithm [21] with mean field-like approximation [22, 23, 24]. Let $\tilde{\mathbf{W}}$ denote a configuration, the joint distribution $p(\mathbf{W} \mid \mathbf{\Phi})$ can be estimated by

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

where $\tilde{\mathcal{N}}(w_{gc})$ represents the neighbors, the nodes that are directly connected to this gene g in cell type c, of w_{gc} corresponding to the fixed configuration $\tilde{\mathbf{W}}$. The complete log likelihood is

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

The Q function in the EM algorithm [21] is

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

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 by Equation (1);
2. Expectation-step: Let $\hat{\mathbf{\Phi}}^{(k)}$ be the estimated parameters in the kth iteration. The Q function $Q(\mathbf{\Phi} \mid \hat{\mathbf{\Phi}}^{(r)})$ can be calculated from the fixed configuration $\tilde{\mathbf{W}}$;
3. Maximization-step: Update $\mathbf{\Phi}$ with $\hat{\mathbf{\Phi}}^{(k+1)}$, which maximizes $Q(\mathbf{\Phi} \mid \hat{\mathbf{\Phi}}^{(k)})$;
4. Obtain the updated $\tilde{\mathbf{W}}$ by sequentially updating $\tilde{w}_{gc}$ by the Gibbs sampler with posterior probability proportional to

$$p(z_{gc} \mid \tilde{w}_{gc})\, p(\tilde{w}_{gc} \mid \mathcal{N}(\tilde{w}_{gc}); \hat{\mathbf{\Phi}}^{(k+1)})$$

5. Repeat steps 2 – 4 until convergence.

Detecting Differentially Expressed Genes

After we obtain the model parameters from the EM algorithm, we use a Gibbs sampler to sample the posterior probabilities. We then use the posterior probability-based definition of FDR [43, 44] to identify DEGs. We first estimate the posterior local FDR p_{gc} using Gibbs sampler,

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

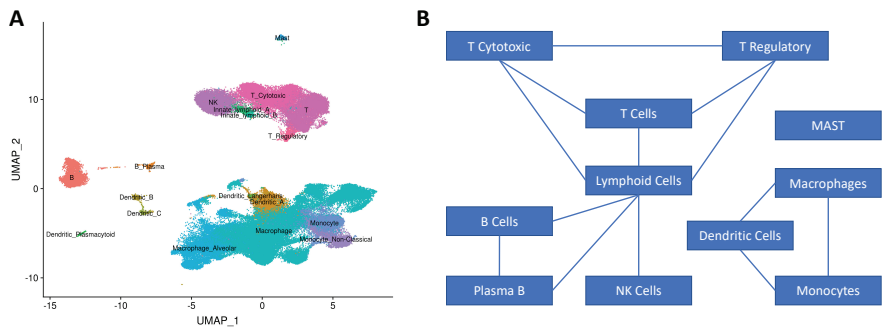

Figure 1 Cell Type Clusters. **(A)** shows the UMAP of eighteen immune cell types, and **(B)** shows the network among these immune cell types that are determined by domain knowledge.

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.

IPF scRNA-seq Data Analysis

The gene expression levels were profiled in each single cell in the IPF scRNA-seq dataset. We used R package `Seurat` [45] to perform data preprocessing 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. After quality control, there were 18,150 protein-coding genes profiled in 114, 364 cells. 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. Each single cell corresponds to a cell type. Since about 87% of the cells were myeloid and lymphoid cells, we focused on the immune cells in further analyses. We used UMAP [46], which is a manifold learning technique for dimension reduction, to plot the cell types. 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 [45, 47] 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 [48] and IntAct [49]. For these 2,000 highly variable genes, the BioGrid database have 5,400 edges, while there are 3,104 edges in the IntAct database. The two gene networks are visualized in Supplementary Figure 1 [see Additional file 1]. There is an overlap of 1,754 edges between the two databases. In addition, the dependencies among cells were determined by domain knowledge (Figure 1B).

For these 2,000 genes across 18 cell types, we fitted two separate MRF models utilizing gene networks from BioGrid and IntAct, respectively. 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. These were the two main models of our analysis, and we labeled them as Main model with BioGrid and Main model with IntAct, respectively. In addition, the model that used only two-sample t-test statistics (no gene and cell networks) was labeled as Main model, which was not an MRF model.

In addition, we considered two sets of supplementary models. First, in order to assess the importance of cell network structures in the MRF model, we considered two additional cell networks C1 and C2 (Supplementary Figure 3) [see Additional file 1]. We used the same observed DE evidence as well as the gene network structures in the Main models, and fitted four additional MRF models with cell networks C1 and C2. We labeled these models with the additional annotation of C1 and C2, respectively. Second, our method can be generalized to use observed DE evidence from other existing DE methods in addition to the two-sample t-test. Most DE methods for scRNA-seq data typically output p-values and log-fold changes for their differential expression analysis. Our model can be readily applied to these existing models with ease. We chose the following two methods as examples: Wilcoxon and `MAST`. Based on the simulation results, these two methods generally have stable and robust performance, especially when the number of cells in each subject is small. In our IPF scRNA-seq data, many cell types have a limited number of cells per subject, i.e., T Regulatory cells and B Plasma cells, so these two methods are more suitable in our IPF analysis. We used `Seurat` to obtain DE results for these two models. Then we transformed the p-values to z-scores and used the sign of log-fold changes to determine the sign of z-scores. With an additional set of observed DE evidence (z-scores), we used the same gene sets and biological networks in the Main models to fit four additional MRF models. We labeled these models accordingly. For instance, the model that only considered DE results from `MAST` analysis was labeled as `MAST` model, and the MRF model that incorporated `MAST` DE results with BioGrid gene network was labeled as `MAST` with BioGrid. A table summarizing the models we used in this manuscript is provided in Supplementary Table 1 [see Additional File 1].

Simulation Study

Simulation study was conducted to assess the performance of our proposed MRF procedure. The single cell RNA-seq data are count data with zero inflation due to dropout and over-dispersion; thus, a zero-inflated negative binomial (ZINB) distribution is suitable to model the read counts and excessive of zeros in the scRNA-seq data. The zero-inflated negative binomial distribution consists of three parameters: mean, dispersion, and inflation. We obtained these parameters based on the IPF scRNA-seq data. In details, for each cell type, we fitted a ZINB distribution across $2,000$ highly variable genes. We obtained three vectors of estimated parameters for the mean, dispersion, and inflation of length $2,000$. Then we fitted a gamma distribution for the estimated means, a gamma distribution for the estimated dispersion parameters, and a beta distribution for the estimated inflation parameters.

We repeated this for eighteen cell types; thus, for each cell type, we had an estimated gamma distribution for the mean, an estimated gamma distribution for the dispersion, and an estimated beta distribution for the inflation.

For our simulated data, the number of genes was set at $1,000$, and we set the number of cell types to be 18, which was the same as that in the IPF scRNA-seq data. Two groups were considered. We set the number of subjects in each group to be 15, and the number of cells for each cell type in each subject to be 50. We also varied the number of subjects and number of cells in the sub-settings. The dimension of our simulated data thus was `# genes` $\times$ `# cell types` $\times$ `# groups` $\times$ `# subjects per group` $\times$ `# cells per cell type per subject`. For network structures, the cell network was set to be the same as that in the IPF scRNA-seq analysis (Main models). For gene network, we randomly selected η percentage genes to be connected. We varied $\eta = 0.2, 0.4, 0.6, 0.8$ to reflect different proportions of connections in the gene network. For differential expression, we first randomly selected κ percentage genes to be DE, which gave us the initial states. With the initial states, we then used the gene and cell network structures to get the latent states by a Gibbs sampler. In each round of Gibbs sampling, the latent states were updated according to Equation (2). We varied $\kappa = 0.2, 0.4$ to reflect different proportions of DEGs in each setting. Then we simulated the expression count data with the zero-inflated negative binomial distribution with mean μ_c, dispersion ϕ_c, and inflation γ_c. The three parameters were sampled from the fitted gamma and beta distributions as mentioned before, and they were cell-type specific and subject-specific. For each individual in the control group, the expression data for each gene were generated from $\text{ZINB}\,(\mu_c, \phi_c, \gamma_c)$. For the case group, if the latent state was 0, the count data were generated from $\text{ZINB}\,(\mu_c, \phi_c, \gamma_c)$; if the latent state was 1, the count data were generated from $\text{ZINB}\,(\tau \cdot \mu_c, \phi_c, \gamma_c)$, where $\tau = \lambda$ and $1/\lambda$ with equal probability. We chose λ to be 2. We performed the same preprocessing and quality control procedures as with the IPF single cell analysis. Then we used test statistics from two-sample t-test as the input of observed DE evidence, and incorporated the above simulated gene/cell networks to construct the MRF model. In order to reflect the role of the number of subjects or cells in our simulations, we considered two sub-scenarios by varying the number of subjects in each group (case/control) to be 30 (Scenario A) and the number of cells per cell type in each subject to be 100 (Scenario B). In addition, we also considered the case when λ is 3 (Scenario C). For each setting, we repeated the simulation 100 times.

We compared the results from our proposed MRF model with nine other methods: the two-sample t-test; the Wilcoxon Rank Sum test; the likelihood ratio test that adopts a logistic regression framework (LR); `MAST`; three pseudo-bulk methods: `DESeq2`, `edgeR`, and `limma`; and two mixed model methods, `dream` and its updated version `dream2`. `dream` and `dream2` [50] were originally designed for bulk RNA-seq studies with repeated measurements and were used in the comparative analysis by Crowell et al. [10]. The original version of `dream` uses `voom`'s [51] precision weights. In its updated version, `dream2`, it adopts the new weighting scheme in the `variancePartition` [52]. We chose these methods because they are representative and have shown fairly robust performance as noted in the review papers [8, 9, 10]. We believe that this was sufficient to demonstrate the performance of our method.

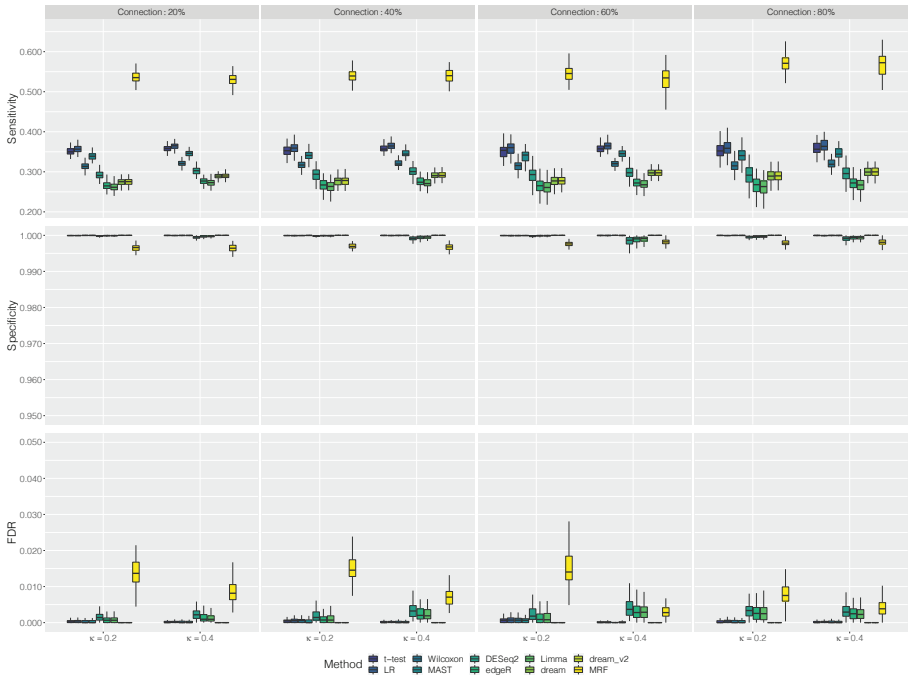

Figure 2 Simulation Results. The results are plotted in terms of sensitivity, specificity and FDR for two-sample t-test, the likelihood ratio test that adopts a logistic regression framework, Wilcoxon rank sum test, `MAST`, three pseudo-bulk methods: `DESeq2`, `edgeR`, and `limma`, two mixed model methods: `dream` and its updated version `dream2`, and the proposed MRF model. Each box-plot represents 100 replications.

Results

Simulation Study

The simulation results are shown in Figure 2, and the results of the three sub-settings (Scenarios A, B and C) are shown in Supplementary Figure 2A-C [see Additional file 1]. The adjusted p-value was set at 0.05. Sensitivity is the fraction of DEGs correctly identified by the method; specificity is the fraction of EE genes identified correctly; and FDR is calculated by the ratio of number of false positives to the number of DEGs identified. Each box-plot represents 100 replications. We note that our proposed MRF model performed significantly better than the other nine methods in terms of sensitivity. The traditional methods have fairly robust performance. The three pseudo-bulk methods have slightly lower specificity and higher FDR compared with other methods. From the two sub-settings, we note that when the number of subjects or number of cells per cell type increases, the performance of the pseudo-bulk methods and methods based on mixed models also increases as expected. Our proposed MRF model still outperforms under these sub-settings. Methods like `MAST` and Wilcoxon still have fairly robust performance, especially when the number of cells in each subject is small. In addition, in order to assess the impact of different thresholds of adjusted p-values on sensitivity, specificity, and FDR, and to see the trade-off directly, we chose three other adjusted p-value cutoffs: 0.01, 0.1, and 0.2 in addition to 0.05. In Supplementary Figure 2D [see Additional file 1], we plot sensitivity, specificity, and FDR for our proposed MRF model for the eight cases (corresponding to Figure 2) under different adjusted p-value cutoffs.

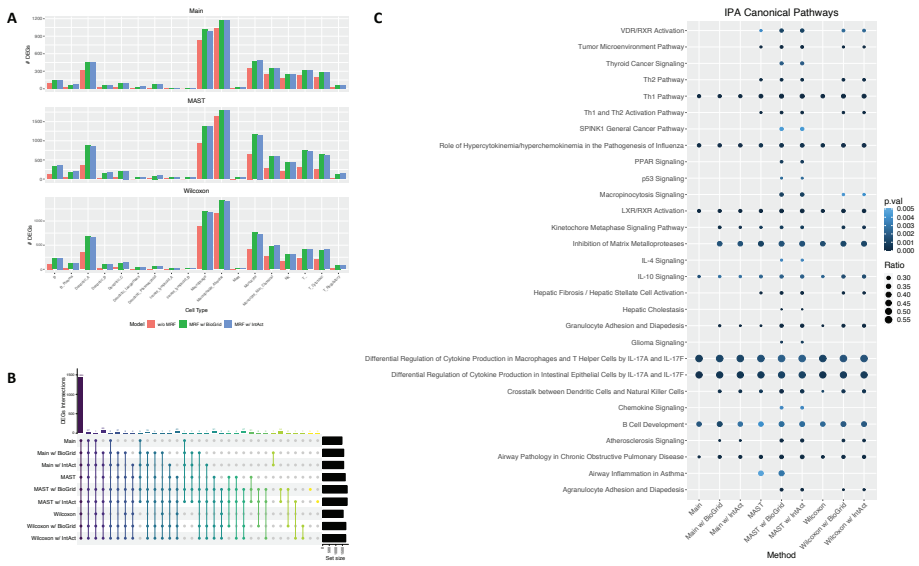

Figure 3 Results from IPF scRNA-seq Data Analysis. **(A)** Number of cell-type specific DEGs identified using original test statistics alone, and two MRF models with the BioGrid and IntAct database for four types of test statistic inputs: two-sample t-test, the Wilcoxon rank sum test, and `MAST`. For instance, the upper-left figure in **(A)** shows the number of cell-type specific DEGs inferred using two-sample t-test alone (w/o MRF), and the DE results from two MRF models that incorporated t-test statistics as observed DE evidence and two gene network structures (MRF w/ BioGrid and MRF w/ IntAct, respectively). **(B)** The UpSet plot shows the overlap of DEGs identified by the original models and our proposed MRF models utilizing the BioGrid and IntAct gene networks. **(C)** The top canonical pathways from Ingenuity Pathways Analysis. The p-value cutoff was set at 0.005 and gene ratio cutoff was set at 0.25 in order to better visualize top enriched pathways.

We note that when the adjusted p-value threshold increases, sensitivity increases and specificity decreases for all eight cases as expected. Our proposed MRF model achieved the desired FDR control.

IPF scRNA-seq Data Analysis

For the Main model with the BioGrid gene network, the estimated parameters were $\gamma = -0.33, \beta_{\text{gene}} = 0.18$, and $\beta_{\text{cell}} = 0.22$, whereas for the Main model with the IntAct gene network, the estimated model parameters were $\gamma = -0.38, \beta_{\text{gene}} = 0.26$, and $\beta_{\text{cell}} = 0.22$. We note that β_{gene} and β_{cell} were comparable here in both models. For DE analysis, we set the significance level at $\alpha = 0.01$ and the corresponding posterior probability cutoff was around 0.91 for both models. Out of 2,000 genes across 18 cell types, the Main MRF model with BioGrid gene network identified 1,605 genes that were found DE in at least one cell type. For the IntAct gene network, the Main MRF model identified 1,601 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 identified 1,472 DEGs. In addition to t-test statistics as input for observed DE evidence, we listed the number of DEGs identified by the Wilcoxon test, and `MAST`, and their corresponding MRF models in Table 1. The parameter estimates for these additional models were shown in Supplementary Figure 4 [see Additional file 1]. We compared cell-type specific DEGs

Table 1 Number of DEGs Identified Under Three Model Settings with Three Types of Test Statistics

Input Statistics	Models	# DEGs[a]	# Cell-type Specific DEGs[b]
t-test	Main	1,472	3,607
	Main w/ BioGrid	1,605	4,880
	Main w/ IntAct	1,601	4,875
MAST	MAST	1,721	4,826
	MAST w/ BioGrid	1,870	8,835
	MAST w/ IntAct	1,867	8,809
Wilcoxon	Wilcoxon	1,562	3,945
	Wilcoxon w/ BioGrid	1,767	6,526
	Wilcoxon w/ IntAct	1,759	6,474

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

inferred by the model using the original test statistics alone, and two MRF models with the BioGrid and IntAct databases for three types of test statistic inputs: the Student's t-test, the Wilcoxon rank sum test, and `MAST` in Figure 3A. We also compared DEGs identified across three types of test statistics. The UpSet [53] plot (Figure 3B) shows the overlap of DEGs identified across different models. We discovered that a major proportion of genes were found DE in all models. The detailed gene lists are provided in the Supplementary Excel file [see Additional file 3].

For different test statistics as observed DE evidence, our model was able to identify an additional set of novel DEGs utilizing gene-gene and cell-cell networks. We performed pathway analysis using Ingenuity Pathways Analysis (IPA, QIAGEN Inc.) [54] to get better biological insights with the inferred DEGs. In order to better visualize top enriched pathways, the p-value cutoff was set at 0.005 and gene ratio cutoff was set at 0.25. The complete list of pathways enriched by each model is provided in Supplementary Excel File [see Additional file 4]. Based on the IPA results in Figure 3C, we found that most pathways were enriched in all twelve models, and our proposed MRF models were able to identify an additional set of pathways that were related to IPF. In fact, we saw that a number of pathways that were related to the T helper cells were found significantly associated with IPF, which was not a surprise based on previous research [55, 56, 57, 58, 59]. In particular, three pathways: Th1 Pathway, Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL-17A and IL-17F pathway, and Differential Regulation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F were found to be significantly associated with IPF under all models. Previous studies [55, 60, 61] showed that IL-17, the cytokines produced by the Th17 cells, participated in the development and progression of pulmonary fibrosis diseases. In addition, our proposed MRF model was able to identify the additional Th2 Pathway, and Th1 and Th2 Activation Pathway to be statistically significant. Previous findings [58, 62] showed that the Th2 cells stimulated fibroblast proliferation and activation, and promoted pulmonary fibrosis. In fact, Th2 responses were related to a number of pulmonary diseases. In addition, the VDR/RXR Activation pathway was also found significant in our MRF models, which was previously found significantly associated

with IPF by Boon et al [63]. The authors also noted that in a mouse study [64], the VDR-deficient mice failed to develop experimental allergic asthma, and this suggested that vitamin D play a key role in the generation of Th2-driven inflammation in lung diseases. Moreover, the MRF model also identified the hepatic fibrosis and hepatic stellate cell activation pathway. Tsuchida et al. [65] noted in their paper that this pathway was well-established as the central driver of hepatic fibrosis in human liver. In addition, the authors discovered that HSC-specific deletion of integrin αv protects mice from hepatic, renal and pulmonary fibrosis. To sum up, the inferred canonical pathways from our approach have biological meanings and are strongly related to IPF.

Furthermore, Supplementary Figures 5 and 6 show the DE analysis results with respect to different cell networks when we fixed the observed DE evidence and gene networks as the same in the Main models. With different cell network structures, the MRF models yielded comparable parameter estimates (Supplementary Figure 5). The UpSet plot in Supplementary Figure 6 shows the overlap of DEGs identified with three cell network structures. We note that the DE results were consistent across different cell networks with slight variations. In addition, the IPA enrichment results also demonstrate that our MRF models have fairly robust performance with respect to different cell network structures.

Conclusions and Discussions

By borrowing information through known biological networks, our proposed method, MRFscRNAseq, provides differential expression analysis to identify cell-type specific DEGs for scRNA-seq data sets with increased statistical power. With observed DE evidence, it utilizes a Markov Random Field model to appropriately take gene network information as well as dependencies among cell types into account. We implemented an Expectation-Maximization (EM) algorithm with mean field-like approximation to estimate model parameters and a Gibbs sampler to infer DE status. Simulation study showed that our method has better power to detect cell-type specific DEGs than conventional methods while appropriately controlling type I error rate. In the differential expression analysis using an IPF scRNA-seq data set, we have showed that our method is able to identify an additional set of novel DEGs using summary statistics from various existing differential expression methods. Pathway analysis with IPA also suggests that the additional set of pathways have biological meanings that are strongly correlated with IPF.

For gene networks, we utilized two protein-protein interaction network databases in the IPF scRNA-seq data application, BioGrid and IntAct. In fact, our method can be adapted to other networks that have similar structures as BioGrid or IntAct PPIN, i.e., KEGG pathways [66]. Furthermore, our model can be readily extended to many other existing DE methods with ease, just like what we have done with Wilcoxon test or `MAST` in the IPF scRNA-seq data analysis.

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 DEGs 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.

Supplementary Information

Abbreviations
DE: Differential expression; DEGs: Differentially expressed genes; MRF: Markov Random Field; EM: Expectation-Maximization; IPF: Idiopathic pulmonary fibrosis; scRNA-seq: single-cell RNA-sequencing. KEGG: Kyoto Encyclopedia of Genes and Genomes; GWAS: Genome-wide association studies; PPIN: Protein-protein interaction network; MAST: Model-based Analysis of Single-cell Transcriptomics.

Authors' contributions
HL and HZ designed the method and wrote the manuscript. HL implemented the algorithm and wrote the R package. BZ helped with the simulation analysis. ZX helped with the IPF scRNA-seq analysis. TA and NK provided the IPF scRNA-seq data set and revised the manuscript. All authors read and approved the final manuscript.

Funding
Supported in part by NIH grant GM134005 and NSF grant DMS 1902903 (HZ). NIH NHLBI grants R01HL127349, R01HL141852, U01HL145567, UH2HL123886 to NK, and a generous gift from Three Lakes Partners to NK.

Availability of data and materials
The IPF scRNA-seq data set is available in the Gene Expression Omnibus repository (GSE136831) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136831). The R package MRFscRNAseq and supplementary results is available on GitHub (https://github.com/eddiehli/MRFscRNAseq).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
NK served as a consultant to Biogen Idec, Boehringer Ingelheim, Third Rock, Pliant, Samumed, NuMedii, Theravance, LifeMax, Three Lake Partners, Optikira, Astra Zeneca, Veracyte, Augmanity and CSL Behring, over the last 3 years, reports Equity in Pliant and a grant from Veracyte, Boehringer Ingelheim, BMS and non-financial support from MiRagen and Astra Zeneca. NK has IP on novel biomarkers and therapeutics in IPF licensed to Biotech. The authors declare that they have no other competing interests.

Acknowledgements
Not applicable.

Author details
[1] Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06511, USA. [2] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA. [3] Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT 06520, USA.

References
1. Kharchenko, P.V., Silberstein, L., Scadden, D.T.: Bayesian approach to single-cell differential expression analysis. Nature Methods **11**(7), 740–742 (2014). doi:10.1038/nmeth.2967. Accessed 2020-01-28
2. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S., Gottardo, R.: MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biology **16**(1), 278 (2015). doi:10.1186/s13059-015-0844-5. Accessed 2020-01-28
3. Korthauer, K.D., Chu, L.-F., Newton, M.A., Li, Y., Thomson, J., Stewart, R., Kendziorski, C.: A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology **17**(1), 222 (2016). doi:10.1186/s13059-016-1077-y. Accessed 2020-01-28
4. Miao, Z., Deng, K., Wang, X., Zhang, X.: DEsingle for detecting three types of differential expression in single-cell RNA-seq data **34**(18), 3223–3224. doi:10.1093/bioinformatics/bty332. Accessed 2020-01-28
5. Wang, T., Nabavi, S.: SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data **145**, 25–32. doi:10.1016/j.ymeth.2018.04.017. Accessed 2020-01-28
6. Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S., Beck, A.H.: EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics **32**(4), 533–541 (2016). doi:10.1093/bioinformatics/btv634. Accessed 2020-01-28
7. Delmans, M., Hemberg, M.: Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics **17**(1), 110 (2016). doi:10.1186/s12859-016-0944-6. Accessed 2020-01-28
8. Soneson, C., Robinson, M.D.: Bias, robustness and scalability in single-cell differential expression analysis **15**(4), 255–261. doi:10.1038/nmeth.4612. Accessed 2021-03-22
9. Wang, T., Li, B., Nelson, C.E., Nabavi, S.: Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data **20**(1), 40. doi:10.1186/s12859-019-2599-6. Accessed 2021-03-22

10. Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., Robinson, M.D.: muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data **11**(1), 6077. doi:10.1038/s41467-020-19894-4. Accessed 2021-03-22
11. Zimmerman, K.D., Espeland, M.A., Langefeld, C.D.: A practical solution to pseudoreplication bias in single-cell studies **12**(1), 738. doi:10.1038/s41467-021-21038-1. Accessed 2021-03-22
12. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., Levine, A.J., La Manno, G., Skinnider, M.A., Courtine, G.: Confronting false discoveries in single-cell differential expression. preprint, Bioinformatics (March 2021). doi:10.1101/2021.03.12.435024. http://biorxiv.org/lookup/doi/10.1101/2021.03.12.435024 Accessed 2021-06-23
13. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data **26**(1), 139–140. doi:10.1093/bioinformatics/btp616. Accessed 2020-01-28
14. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 **15**(12), 550. doi:10.1186/s13059-014-0550-8. Accessed 2020-01-28
15. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**(7), 47–47 (2015). doi:10.1093/nar/gkv007. Accessed 2021-06-23
16. Wei, Z., Li, H.: A Markov random field model for network-based analysis of genomic data. Bioinformatics **23**(12), 1537–1544 (2007). doi:10.1093/bioinformatics/btm129. Accessed 2020-01-28
17. Wei, Z., Li, H.: 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 (2008). doi:10.1214/07–AOAS145. Accessed 2020-01-28
18. Chen, M., Cho, J., Zhao, H.: Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. PLoS Genetics **7**(4), 1001353 (2011). doi:10.1371/journal.pgen.1001353. Accessed 2020-03-31
19. Lin, Z., Sanders, S.J., Li, M., Sestan, N., State, M.W., Zhao, H.: 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 (2015). doi:10.1214/14-AOAS802. Accessed 2020-01-28
20. Efron, B.: Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association **99**(465), 96–104 (2004). doi:10.2307/27590356
21. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the *EM* algorithm **39**(1), 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x. Accessed 2021-03-22
22. Zhang, J.: The mean field theory in EM procedures for Markov random fields. IEEE Transactions on Signal Processing **40**(10), 2570–2583 (1992). doi:10.1109/78.157297. Accessed 2020-04-01
23. Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognition **36**(1), 131–144 (2003). doi:10.1016/S0031-3203(02)00027-4. Accessed 2020-04-01
24. Lin, Z., Li, M., Sestan, N., Zhao, H.: 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) (2016). doi:10.1515/sagmb-2015-0070. Accessed 2020-01-28
25. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G., Raby, B.A., Deluliis, G., Januszyk, M., Duan, Q., Arnett, H.A., Siddiqui, A., Washko, G.R., Homer, R., Yan, X., Rosas, I.O., Kaminski, N.: Single-cell rna-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Science Advances **6**(28) (2020). doi:10.1126/sciadv.aba1983
26. Martinez, F.J., Collard, H.R., Pardo, A., Raghu, G., Richeldi, L., Selman, M., Swigris, J.J., Taniguchi, H., Wells, A.U.: Idiopathic pulmonary fibrosis. Nature Reviews Disease Primers **3**(1), 17074 (2017). doi:10.1038/nrdp.2017.74. Accessed 2020-01-28
27. Barratt, S., Creamer, A., Hayton, C., Chaudhuri, N.: Idiopathic Pulmonary Fibrosis (IPF): An Overview. Journal of Clinical Medicine **7**(8), 201 (2018). doi:10.3390/jcm7080201. Accessed 2020-01-28
28. Ley, B., Collard, H.R., King, T.E.: Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine **183**(4), 431–440 (2011). doi:10.1164/rccm.201006-0894CI. Accessed 2020-03-31
29. Raghu, G., Chen, S.-Y., Yeh, W.-S., Maroni, B., Li, Q., Lee, Y.-C., Collard, H.R.: 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 (2014). doi:10.1016/S2213-2600(14)70101-8. Accessed 2020-01-28
30. Fingerlin, T.E., Murphy, E., Zhang, W., Peljto, A.L., Brown, K.K., Steele, M.P., Loyd, J.E., Cosgrove, G.P., Lynch, D., Groshong, S., Collard, H.R., Wolters, P.J., Bradford, W.Z., Kossen, K., Seiwert, S.D., du Bois, R.M., Garcia, C.K., Devine, M.S., Gudmundsson, G., Isaksson, H.J., Kaminski, N., Zhang, Y., Gibson, K.F., Lancaster, L.H., Cogan, J.D., Mason, W.R., Maher, T.M., Molyneaux, P.L., Wells, A.U., Moffatt, M.F., Selman, M., Pardo, A., Kim, D.S., Crapo, J.D., Make, B.J., Regan, E.A., Walek, D.S., Daniel, J.J., Kamatani, Y., Zelenika, D., Smith, K., McKean, D., Pedersen, B.S., Talbert, J., Kidd, R.N., Markin, C.R., Beckman, K.B., Lathrop, M., Schwarz, M.I., Schwartz, D.A.: Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nature Genetics **45**(6), 613–620 (2013). doi:10.1038/ng.2609. Accessed 2020-03-31
31. Noth, I., Zhang, Y., Ma, S.-F., Flores, C., Barber, M., Huang, Y., Broderick, S.M., Wade, M.S., Hysi, P., Scuirba, J., Richards, T.J., Juan-Guardela, B.M., Vij, R., Han, M.K., Martinez, F.J., Kossen, K., Seiwert, S.D., Christie, J.D., Nicolae, D., Kaminski, N., Garcia, J.G.: Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. The Lancet Respiratory Medicine **1**(4), 309–317 (2013). doi:10.1016/S2213-2600(13)70045-6. Accessed 2020-03-31
32. Allen, R.J., Porte, J., Braybrooke, R., Flores, C., Fingerlin, T.E., Oldham, J.M., Guillen-Guio, B., Ma, S.-F., Okamoto, T., John, A.E., Obeidat, M., Yang, I.V., Henry, A., Hubbard, R.B., Navaratnam, V., Saini, G., Thompson, N., Booth, H.L., Hart, S.P., Hill, M.R., Hirani, N., Maher, T.M., McAnulty, R.J., Millar, A.B.,

Molyneaux, P.L., Parfrey, H., Rassl, D.M., Whyte, M.K.B., Fahy, W.A., Marshall, R.P., Oballa, E., Bossé, Y., Nickle, D.C., Sin, D.D., Timens, W., Shrine, N., Sayers, I., Hall, I.P., Noth, I., Schwartz, D.A., Tobin, M.D., Wain, L.V., Jenkins, R.G.: 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 (2017). doi:10.1016/S2213-2600(17)30387-9. Accessed 2020-03-31

33. Allen, R.J., Guillen-Guio, B., Oldham, J.M., Ma, S.-F., Dressen, A., Paynton, M.L., Kraven, L.M., Obeidat, M., Li, X., Ng, M., Braybrooke, R., Molina-Molina, M., Hobbs, B.D., Putman, R.K., Sakornsakolpat, P., Booth, H.L., Fahy, W.A., Hart, S.P., Hill, M.R., Hirani, N., Hubbard, R.B., McAnulty, R.J., Millar, A.B., Navaratnam, V., Oballa, E., Parfrey, H., Saini, G., Whyte, M.K.B., Zhang, Y., Kaminski, N., Adegunsoye, A., Strek, M.E., Neighbors, M., Sheng, X.R., Gudmundsson, G., Gudnason, V., Hatabu, H., Lederer, D.J., Manichaikul, A., Newell, J.D., O'Connor, G.T., Ortega, V.E., Xu, H., Fingerlin, T.E., Bossé, Y., Hao, K., Joubert, P., Nickle, D.C., Sin, D.D., Timens, W., Furniss, D., Morris, A.P., Zondervan, K.T., Hall, I.P., Sayers, I., Tobin, M.D., Maher, T.M., Cho, M.H., Hunninghake, G.M., Schwartz, D.A., Yaspan, B.L., Molyneaux, P.L., Flores, C., Noth, I., Jenkins, R.G., Wain, L.V.: Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine **201**(5), 564–574 (2020). doi:10.1164/rccm.201905-1017OC. Accessed 2020-03-31
34. Zuo, F., Kaminski, N., Eugui, E., Allard, J., Yakhini, Z., Ben-Dor, A., Lollini, L., Morris, D., Kim, Y., DeLustro, B., Sheppard, D., Pardo, A., Selman, M., Heller, R.A.: 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 (2002). doi:10.1073/pnas.092134099. Accessed 2020-11-08
35. Meltzer, E.B., Barry, W.T., D'Amico, T.A., Davis, R.D., Lin, S.S., Onaitis, M.W., Morrison, L.D., Sporn, T.A., Steele, M.P., Noble, P.W.: Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. BMC Medical Genomics **4**(1), 70 (2011). doi:10.1186/1755-8794-4-70. Accessed 2020-03-31
36. Yang, I.V., Coldren, C.D., Leach, S.M., Seibold, M.A., Murphy, E., Lin, J., Rosen, R., Neidermyer, A.J., McKean, D.F., Groshong, S.D., Cool, C., Cosgrove, G.P., Lynch, D.A., Brown, K.K., Schwarz, M.I., Fingerlin, T.E., Schwartz, D.A.: Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. Thorax **68**(12), 1114–1121 (2013). doi:10.1136/thoraxjnl-2012-202943. Accessed 2020-03-31
37. Yue, X., Lu, J., Auduong, L., Sides, M.D., Lasky, J.A.: Overexpression of Sulf2 in idiopathic pulmonary fibrosis. Glycobiology **23**(6), 709–719 (2013). doi:10.1093/glycob/cwt010. Accessed 2020-03-31
38. Deng, N., Sanchez, C.G., Lasky, J.A., Zhu, D.: Detecting Splicing Variants in Idiopathic Pulmonary Fibrosis from Non-Differentially Expressed Genes. PLoS ONE **8**(7), 68352 (2013). doi:10.1371/journal.pone.0068352. Accessed 2020-03-31
39. Nance, T., Smith, K.S., Anaya, V., Richardson, R., Ho, L., Pala, M., Mostafavi, S., Battle, A., Feghali-Bostwick, C., Rosen, G., Montgomery, S.B.: Transcriptome Analysis Reveals Differential Splicing Events in IPF Lung Tissue. PLoS ONE **9**(3), 92111 (2014). doi:10.1371/journal.pone.0092111. Accessed 2020-03-31
40. McDonough, J.E., Ahangari, F., Li, Q., Jain, S., Verleden, S.E., Herazo-Maya, J., Vukmirovic, M., DeIuliis, G., Tzouvelekis, A., Tanabe, N., Chu, F., Yan, X., Verschakelen, J., Homer, R.J., Manatakis, D.V., Zhang, J., Ding, J., Maes, K., De Sadeleer, L., Vos, R., Neyrinck, A., Benos, P.V., Bar-Joseph, Z., Tantin, D., Hogg, J.C., Vanaudenaerde, B.M., Wuyts, W.A., Kaminski, N.: Transcriptional regulatory model of fibrosis progression in the human lung. JCI Insight **4**(22), 131597 (2019). doi:10.1172/jci.insight.131597. Accessed 2020-11-08
41. Vukmirovic, M., Kaminski, N.: Impact of Transcriptomics on Our Understanding of Pulmonary Fibrosis. Frontiers in Medicine **5**, 87 (2018). doi:10.3389/fmed.2018.00087. Accessed 2020-11-08
42. Besag, J.: On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society: Series B (Methodological) **48**(3), 259–279 (1986). doi:10.1111/j.2517-6161.1986.tb01412.x. Accessed 2020-04-01
43. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W.: On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. Journal of Computational Biology **8**(1), 37–52 (2001). doi:10.1089/106652701300099074
44. Li, H., Wei, Z., Maris, J.: A hidden Markov random field model for genome-wide association studies. Biostatistics **11**(1), 139–150 (2010). doi:10.1093/biostatistics/kxp043. Accessed 2020-04-01
45. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology **36**(5), 411–420 (2018). doi:10.1038/nbt.4096. Accessed 2020-01-28
46. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. 1802.03426. Accessed 2021-03-22
47. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G.: Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods **10**(11), 1093–1095 (2013). doi:10.1038/nmeth.2645. Accessed 2020-09-15
48. Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2019 update. Nucleic Acids Research **47**(D1), 529–541 (2019). doi:10.1093/nar/gky1079. Accessed 2020-09-15
49. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., Hermjakob, H.: The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research **42**(D1), 358–363 (2014). doi:10.1093/nar/gkt1115. Accessed 2020-09-15
50. Hoffman, G.E., Roussos, P.: Dream: powerful differential expression analysis for repeated measures designs. Bioinformatics **37**(2), 192–201 (2021). doi:10.1093/bioinformatics/btaa687. Accessed 2021-06-23
51. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: voom: precision weights unlock linear model analysis tools for

RNA-seq read counts. Genome Biology **15**(2), 29 (2014). doi:10.1186/gb-2014-15-2-r29. Accessed 2021-06-23
52. Hoffman, G.E., Schadt, E.E.: variancePartition: interpreting drivers of variation in complex gene expression studies. BMC Bioinformatics **17**(1), 483 (2016). doi:10.1186/s12859-016-1323-z. Accessed 2021-06-23
53. Conway, J.R., Lex, A., Gehlenborg, N.: UpSetR: an r package for the visualization of intersecting sets and their properties **33**(18), 2938–2940. doi:10.1093/bioinformatics/btx364. Accessed 2021-03-22
54. Krämer, A., Green, J., Pollard, J., Tugendreich, S.: Causal analysis approaches in ingenuity pathway analysis **30**(4), 523–530. doi:10.1093/bioinformatics/btt703. Accessed 2021-03-22
55. Mi, S., Li, Z., Yang, H.-Z., Liu, H., Wang, J.-P., Ma, Y.-G., Wang, X.-X., Liu, H.-Z., Sun, W., Hu, Z.-W.: Blocking IL-17a promotes the resolution of pulmonary inflammation and fibrosis via TGF-beta1–dependent and –independent mechanisms **187**(6), 3003–3014. doi:10.4049/jimmunol.1004081. Accessed 2021-03-22
56. Wuyts, W.A., Agostini, C., Antoniou, K.M., Bouros, D., Chambers, R.C., Cottin, V., Egan, J.J., Lambrecht, B.N., Lories, R., Parfrey, H., Prasse, A., Robalo-Cordeiro, C., Verbeken, E., Verschakelen, J.A., Wells, A.U., Verleden, G.M.: The pathogenesis of pulmonary fibrosis: a moving target **41**(5), 1207–1218. doi:10.1183/09031936.00073012. Accessed 2021-03-22
57. Tan, H.-L., Regamey, N., Brown, S., Bush, A., Lloyd, C.M., Davies, J.C.: The th17 pathway in cystic fibrosis lung disease **184**(2), 252–258. doi:10.1164/rccm.201102-0236OC. Accessed 2021-03-22
58. Desai, O., Winkler, J., Minasyan, M., Herzog, E.L.: The role of immune and inflammatory cells in idiopathic pulmonary fibrosis **5**, 43. doi:10.3389/fmed.2018.00043. Accessed 2021-03-22
59. Kolahian, S., Fernandez, I.E., Eickelberg, O., Hartl, D.: Immune mechanisms in pulmonary fibrosis **55**(3), 309–322. doi:10.1165/rcmb.2016-0121TR. Accessed 2021-03-22
60. Gurczynski, S.J., Moore, B.B.: IL-17 in the lung: the good, the bad, and the ugly **314**(1), 6–16. doi:10.1152/ajplung.00344.2017. Accessed 2021-03-22
61. Zhang, J., Wang, D., Wang, L., Wang, S., Roden, A.C., Zhao, H., Li, X., Prakash, Y.S., Matteson, E.L., Tschumperlin, D.J., Vassallo, R.: Profibrotic effect of IL-17a and elevated IL-17ra in idiopathic pulmonary fibrosis and rheumatoid arthritis-associated lung disease support a direct role for IL-17a/IL-17ra in human fibrotic interstitial lung disease **316**(3), 487–497. doi:10.1152/ajplung.00301.2018. Accessed 2021-03-22
62. Barron, L., Wynn, T.A.: Fibrosis is regulated by th2 and th17 responses and by dynamic interactions between fibroblasts and macrophages **300**(5), 723–728. doi:10.1152/ajpgi.00414.2010. Accessed 2021-03-22
63. Boon, K., Bailey, N.W., Yang, J., Steel, M.P., Groshong, S., Kervitsky, D., Brown, K.K., Schwarz, M.I., Schwartz, D.A.: Molecular phenotypes distinguish patients with relatively stable from progressive idiopathic pulmonary fibrosis (IPF) **4**(4), 5134. doi:10.1371/journal.pone.0005134. Accessed 2021-03-22
64. Wittke, A., Weaver, V., Mahon, B.D., August, A., Cantorna, M.T.: Vitamin d receptor-deficient mice fail to develop experimental allergic asthma **173**(5), 3432–3436. doi:10.4049/jimmunol.173.5.3432. Accessed 2021-03-22
65. Tsuchida, T., Friedman, S.L.: Mechanisms of hepatic stellate cell activation **14**(7), 397–411. doi:10.1038/nrgastro.2017.38. Accessed 2021-04-02
66. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes, 4 (2000)