Human Genomics

RESEARCH

Open Access

Bayesian-frequentist hybrid inference framework for single cell RNA-seq analyses

Gang Han[1], Dongyan Yan[2], Zhe Sun[2], Jiyuan Fang[2], Xinyue Chang[2], Lucas Wilson[1] and Yushi Liu[2*]

Abstract

Background Single cell RNA sequencing technology (scRNA-seq) has been proven useful in understanding cell-specific disease mechanisms. However, identifying genes of interest remains a key challenge. Pseudo-bulk methods that pool scRNA-seq counts in the same biological replicates have been commonly used to identify differentially expressed genes. However, such methods may lack power due to the limited sample size of scRNA-seq datasets, which can be prohibitively expensive.

Results Motivated by this, we proposed to use the Bayesian-frequentist hybrid (BFH) framework to increase the power and we showed in simulated scenario, the proposed BFH would be an optimal method when compared with other popular single cell differential expression methods if both FDR and power were considered. As an example, the method was applied to an idiopathic pulmonary fibrosis (IPF) case study.

Conclusion In our IPF example, we demonstrated that with a proper informative prior, the BFH approach identified more genes of interest. Furthermore, these genes were reasonable based on the current knowledge of IPF. Thus, the BFH offers a unique and flexible framework for future scRNA-seq analyses.

Keywords Bayesian-frequentist hybrid inference, Informative prior, Single-cell RNA-seq

Background

Single cell RNA sequencing (scRNA-seq) is a powerful sequencing technology that allows for the profiling of gene expression in individual cells. Traditional bulk RNA sequencing technologies measure the average expression level of all cells in the population, which mask the uniqueness of each cell. In contrast, isolation of cells is an important step in scRNA-seq. It enables the identification of different cell types within complex tissues [36]. As demonstrated in Keren-Shaul et al.'s research, by analyzing immune cell populations in mouse brains, they discovered a novel microglia type associated with neurodegenerative diseases using scRNA-seq [21].

scRNA-seq has the advantage in processing thousands or even millions of single cells simultaneously [48] and has extensive applications across different fields of biology and medical research. By comparing the gene expression level between patients and healthy controls, scRNA-seq can provide important insights into the disease associated genes and pathways. In drug discovery area, it has become an essential tool to identify novel drug targets and to test the efficacy of drugs on specific cell types. For instance, in the study by Wu et al. [42] on diabetic kidney disease (DKD), they generated single cell data from nearly 1 million cells and analyzed the response of a murine DKD model to five treatment approaches. They found that different medications affected different cell types, and combination therapies achieved better outcomes in rescuing DKD-associated transcriptional changes.

*Correspondence:
Yushi Liu
liu_yushi@lilly.com
[1] Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, TX, USA
[2] Eli Lilly and Company, Lilly Corporate Center, 893 Delaware St, Indianapolis, IN 46225, USA

One important question in analyzing scRNA-seq data is the identification of differentially expressed genes (DEG) between groups. Compared to the gene expression data generated from other sequencing technologies, scRNA-seq data have some unique features including overdispersion, sparsity, high proportion of zeros due to dropout events (i.e., scRNA-seq data only captures a small fraction of the transcriptome of each cell), and the hierarchical structure embedded in the data [10]. Early scRNA-seq studies often collect many cells from one or a few individuals. With the rapid advancement in the technology, scientists have started to collect single cell data from multiple individuals. In a multi-individual, multi-condition experiment, other than cell-to-cell variation within each individual, heterogeneity also exists among different conditions, individuals and across different cell types. Those distinctive challenges need to be considered when we explore DEG in scRNA-seq data.

With more and more gene expression data becoming publicly available, many approaches and tools for the differential expression (DE) analysis have been developed for scRNA-seq data. For example, ZINB-WaVE [32] and ZingeR (Van den [37]) assume the expression counts follow a zero-inflated negative binomial (ZINB) distribution and apply Expectation–Maximization (EM) algorithms to estimate model parameters. In contrast, SCDE [22] models the observed abundance using a mixture of the Poisson (dropout component) and negative binomial (amplification component) distribution. These approaches require several distributional assumptions which may fail to be satisfied by real data. MAST [12] applied a hurdle model to simultaneously model the expression rate and mean expression values for a specific gene, then DE testing is performed between two cell populations using likelihood ratio test statistic. Non-parametric approaches were also applied on analyzing scRNA-seq data, such as Wilcoxon signed rank test and ROSeq [16], both of which use test statistics based on ranks. In summary, many single-cell-specific DE methods which apply different strategies, have been developed in recent years. However, some existing approaches are inappropriate for individual level differential expression testing (such as comparison between patients and healthy controls), as the sampling units for these approaches are cells, not individuals [47]. Failing to account for the intrinsic variability of individuals causes a systematic underestimation of the variance of gene expression, compromising the ability to generate biologically accurate results.

Pseudo-bulk methods, which pool the scRNA-seq counts in the same biological replicate, have been developed to address this variability. Squair et al. [34] evaluated the performance across fourteen different DE methods using eighteen datasets and found pseudo-bulk methods outperform other cell-level based DE methods in scRNA-seq data. Murphy and Skene [29] also recommended the use of pseudo-bulk approaches after the simulation analysis from multiple scenarios. Biased inference and highly inflated type 1 error rates were observed when scientists assume cells from the same individual are statistically independent. Zimmerman, Espeland, and Langefeld [49] proposed a generalized linear mixed model that incorporates a random effect for individual, to address the correlation structure from cells within an individual. NEBULA [20] is an efficient negative binomial mixed model accounting for both individual-level and cell-level overdispersions. Another method IDEAS [47] captures the gene expression profile in each individual by a probability distribution and then compares such distributions across two groups of individuals.

Regardless, pseudo-bulk methods could still lack power to detect genes of interest due to the limited sample size. To overcome these limitations, we propose to use a Bayesian-frequentist hybrid (BFH) inference method to analyze the scRNA-seq data at the individual level. The BFH theoretical framework was originally proposed by Yuan [45] and the computation framework based on the EM algorithm and Monte-Carlo Markov Chain was proposed by Han et al. [19]. In BFH, part of the model parameters is frequentist, and others are Bayesian. The goal of BFH is to obtain estimation of both types of parameters and quantify the variation in the estimation. BFH is achieved by maximizing the likelihood function given the Bayesian parameters and simultaneously minimizing the posterior expected loss function given the frequentist parameters. We extended the work of Han et al. [19] using a linear regression model based on normal distribution, where both the frequentist and Bayesian estimators have tractable analytic forms. We also derive the estimation error (or standard error) of the frequentist and Bayesian parameters. With a point estimate and a standard error of an estimator, we can construct confidence intervals of the coefficients, which can also be used to test whether predictors (such as disease group) are significantly associated with gene expression.

Methods

The hybrid inference in existing literature

BFH inference is designed for models that have both frequentist and Bayesian parameters [45]. Suppose the frequentist and Bayesian parameters are θ_A and θ_B, respectively, the data is $\boldsymbol{Y}$, and the prior for θ_B is $\pi(\theta_B)$. Given a decision $d(\boldsymbol{Y})$, a loss function $W(d(\boldsymbol{Y}), \theta_B)$, and the distribution likelihood $f(\boldsymbol{Y}|\theta_A, \theta_B)$, the hybrid estimators of θ_A and θ_B are

$$\left(\hat{\theta}_A, \check{\theta}_B\right) = \arg\inf\sup \int W(d(\boldsymbol{Y}), \theta_B) f(\boldsymbol{Y}|\theta_A, \theta_B)\pi(\theta_B) d\theta_B,$$

where inf and *sup* were taken in the space of $d(\boldsymbol{Y})$ and θ_A respectively so that $\check{\theta}_B$ minimizes the posterior risk given $\hat{\theta}_A$ and $\hat{\theta}_A$ maximizes the likelihood function given $\check{\theta}_B$. The frequentist parameter $\hat{\theta}_A$ is defined (and can be numerically calculated) as integration of the loss function over the posterior distribution. Yuan [45] proved that the hybrid estimator is a consistent estimator, and the standard error of the hybrid estimators can converge to that of the frequentist estimators. As a result, the variance–covariance matrix can be quantified using Fisher information matrix. Han et al. [19] developed an EM computational algorithm to compute $\left(\hat{\theta}_A, \check{\theta}_B\right)$ for any loss function ensuring that the hybrid inference is applicable to general practical problems and different data settings. Han et al. [19] demonstrated, in extensive simulation studies, that the hybrid inference based on the EM algorithm can outperform Bayesian inference and frequentist inference. In this paper we adopt the EM algorithm in Han et al. [19] to make inference. Data, statistical models, and more details about the computation are given in section "Introduction of frequentist, Bayesian, and hybrid inference in linear regression with conjugate priors".

Single-cell RNA sequencing methods comparison using semi-synthetic dataset

Motivation of semi-synthetic data We employed semi-synthetic data derived from actual single-cell RNA sequencing (scRNA-seq) data to assess the power and false discovery rate (FDR) of our proposed method in comparison to other widely used approaches. Inspired by Li et al.'s work [26], where semi-synthetic data was employed to evaluate bulk RNA-seq differential expression (DE) methods, we sought to extend this approach to scRNA-seq. Traditional simulated datasets often struggle to accurately capture the biological signals and intricate correlation structures present in real datasets, leading to challenges in maintaining cellular population heterogeneity [8]. Consequently, analyses based on diverse simulations and packages may yield conflicting conclusions. For instance, Zimmerman et al. [49] observed superior Type I error rate control and power in mixed models compared to pseudo-bulk methods using simulated data. However, Murphy et al., in a different simulation setup, found that pseudo-bulk methods exhibited the lowest Type II error rate among all tested methods, with equal Type I error rates. Recognizing these discrepancies, we advocate for a more realistic evaluation procedure for DE methods.

Semi-synthetic data source and data generation: HypoMap, a compilation of mouse hypothalamus single-cell RNA sequencing (scRNA-seq) data sourced from 18 publications [35], encompasses 100 normal chow mice, yielding a dataset containing190,710 neuron cells. Given the substantial number of subjects and cells within this dataset, it serves as a valuable resource for generating multiple synthetic datasets. As a subset, 55 mice were identified with a minimum of 1,000 cells each, constituting a total of 170,874 neuron cells. Based on this subset, we derived our semi-synthetic datasets.

In our scRNA-seq semi-synthetic scenario, we randomly selected 20 out of 55 mice. Among these, 10 were arbitrarily assigned to the 'disease' group, while the remaining mice served as the 'normal chow' group. It's important to note that in the original data, the 'disease' group was under normal chow conditions. To achieve around 1,000 cells per mouse, we sampled a specific number of cells from each mouse using a Poisson distribution with a mean of 1,000, ensuring an average of 1,000 cells across the 20 mice.

For the generation of true positives (true differentially expressed genes or true DE genes), we initially focused on genes expressed in at least 10% of the cells, amounting to approximately 8,000 genes. Subsequently, we randomly selected 5% of these genes from the 'disease' group and an artificial fixed effect was introduced to these genes by multiplying the counts under the 'disease' condition by a constant of 2. Consequently, these genes represent true positives, while the remaining genes serve as true negatives (non-DE genes). This process was iterated 100 times to generate 100 semi-synthetic datasets.

Differential expression method selection We carefully selected representative approaches from three domains of previously proposed Differential Expression (DE) methods: mixed models, pseudo-bulk methods, and single-cell methods. Our choice for the mixed model approach was NEBULA [20, 27], edgeR [33], and limma-voom [25].

ScRNA-seq and bulk RNA-seq Data source

Lungmap dataset The Lungmap dataset used in this study was from a published human lung tissue study [39]. The cells were clustered by Seurat v3 [7] and annotated to 31 cell types based on canonical lineage-defining markers. Lungmap dataset served as the reference dataset for Hierarchical XGBoost [9] algorithm to obtain the probability for a cell being an alveolar macrophage cell. In addition to output of probabilities indicating the likelihood of a candidate cell belonging to each individual cell type, HierXGB offers an additional capability. HierXGB can provide cell identity directly using a naive Bayesian

approach, assigning the cell type based on the maximum of such predicted probability.

IPF scRNA-seq dataset The idiopathic pulmonary fibrosis (IPF) scRNA-seq dataset was obtained from a previous study on pulmonary fibrosis (PF) disease mechanisms and the corresponding cell types in human lung tissues [17]. This dataset contains over 114,000 cells from 22 donors who had cell observations. Among these 22 donors, 10 are from the control group and 12 from the IPF group. The IPF dataset served as the query data for HierXGB to obtain the probability for a cell being an alveolar macrophage cell.

IPF bulk RNA-seq dataset We also obtained bulk RNA-seq IPF data from human lung tissues [13] in the previous research on the relationship between chronic hypersensitivity pneumonitis and idiopathic pulmonary fibrosis. This bulk IPF dataset contains 18,838 genes from 103 idiopathic IPF samples and 103 unaffected controls samples. For each gene, the mean differences of expression level between the IPF and control groups were calculated using linear regression with the adjustment of age, sex, race, and smoking history. This difference served as the informative prior of the phenotype coefficient.

Data preprocessing of IPF datasets

For scRNA-seq dataset, we selected genes with average expression level across cells greater than 0.1. We removed cells when the number of detected genes was below the lower 2-percentile or with more than 10% of mitochondrial gene expression. For the cell weight calculation, we aligned Lungmap dataset and IPF single cell RNA-seq datasets by their common genes, resulting in 13,988 common genes and 44,294 and 50,383 cells for Lungmap and IPF, respectively. For coefficient estimation of each gene, we matched the IPF scRNA-seq and IPF bulk datasets and obtained 7,886 common genes.

Introduction of frequentist, Bayesian, and hybrid inference in linear regression with conjugate priors

Here, we would like to introduce each of the methods we attempted to identify genes associated with IPF. In linear regression analysis, given the sample size n and the number of regression parameters p, the data can be arranged in Y of dimension $n \times 1$ as the response, and X of dimension $n \times p$ as the design matrix. The regression parameter in the model β can be arranged in a vector of dimension $p \times 1$.

In our analysis the outcome Y is the weighted average of a gene's expression for a particular cell type (e.g., TGF-β1 in alveolar macrophage) at the individual level, $p=1$, and X the design matrix is composed of a vector of 1 s, X_1 the disease group of the individual (control or IPF), and X_2 the predictive probability of each cell belonging to alveolar macrophage averaged per individual with subsequent negative log transformation. The linear model is $Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$, where I is an identity matrix with dimension n by n. So $Y \sim N(X\beta, \sigma^2 I)$. The regression parameters are $\beta = (\beta_0, \beta_1, \beta_2)$, which β_0, β_1 and β_2 are the intercept, regression parameter for disease group, and regression parameter for probability of the cell belonging to alveolar macrophage, respectively. The likelihood value given data $(\boldsymbol{Y}, \boldsymbol{X})$ is

$$P((\boldsymbol{Y}, \boldsymbol{X})|\beta) = \prod_{i=1}^{n} P\left(y_i = N\left(X_i\beta, \sigma^2\right)\right). \tag{1}$$

The conjugate prior of the regression parameter can be written as $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$. Then the posterior distribution can be derived as

$$P(\beta|(\boldsymbol{Y}, \boldsymbol{X})) \propto P((\boldsymbol{Y}, \boldsymbol{X})|\beta)\pi(\beta) \sim \prod_{i=1}^{n} N\left(X_i\beta, \sigma^2 I|y_i\right) \times N(\mu_\beta, \Sigma_\beta) \sim N\left(\mu_\beta^{new}, \Sigma_\beta^{new}\right), \tag{2}$$

where $\mu_\beta^{new} = \left(\Sigma_\beta^{-1} + X^{\mathrm{T}}X\right)^{-1}\left(\Sigma_\beta^{-1}\mu_\beta + X^{\mathrm{T}}Y\right)$ and $\Sigma_\beta^{new} = \left(\Sigma_\beta^{-1} + X^{\mathrm{T}}X\right)^{-1}\sigma^2$. Without the prior $\pi(\beta)$, the frequentist's estimate and its variance of β can be written as $\mu_\beta^F = \left(X^{\mathrm{T}}X\right)^{-1}\left(X^{\mathrm{T}}Y\right)$ and $\Sigma_\beta^F = \left(X^{\mathrm{T}}X\right)^{-1}\sigma^2$, which is the ordinary least square estimate.

Finally, following the EM algorithm in Han et al. [19], the hybrid Bayesian analysis can be written in the following iterative procedures:

[Step 1.] Initialize parameters $(\beta_0, \beta_1, \beta_2)$ from the frequentist estimates as $\left(\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}\right)$, where β_0 is the intercept, and (β_1, β_2) are the slope parameters for disease group X_1 and the predictive probability of each cell belonging to alveolar macrophage averaged per individual with subsequent negative log transformation X_2, respectively.

[Step 2.] Given the current value of frequentist parameters $\left(\beta_0^{(t)}, \beta_2^{(t)}\right)$, generate data $y_i^B = y_i - X_{0,i}\beta_0^{(t)} - X_{2,i}\beta_2^{(t)}$. From the regression model $Y^B = X_1\beta_1$, obtain $\beta_1^{(t+1)}$ as the posterior mean of β_1, given a conjugate (normal) prior of β_1.

[Step 3.] Given $\beta_1^{(t+1)}$, the posterior mean of β_1, generate data $y_i^F = y_i - X_{1,i}\beta_1^{(t+1)}$. From the regression model

$Y^F = X_0\beta_0 + X_2\beta_2$, obtain $\left(\beta_0^{(t+1)}, \beta_2^{(t+1)}\right)$ as frequentist estimate of (β_0, β_2).

[Step 4.] Iterate steps 2–3 as in EM algorithm.

For frequentist, Bayesian, and hybrid inferences can all generate parameter estimates and the corresponding estimation variances. An estimate and its variance are used to construct a 95% confidence interval (estimate minus and plus 1.96 times of the standard error) and to calculate a *p*-value from the two-sided test (by calculating a z-score of estimate divided by the standard error) of whether this value is equal 0 or not based on the underlying normal distribution.

Acquiring weights for alveolar macrophage cells using HierXGB method

Alveolar macrophage cells have been recognized to play a crucial role in the pathogenesis of IPF [2, 46]. Rather than analyzing gene expression levels across all cell types, we are specifically interested in the association between gene expression levels and disease group (IPF vs. control) in alveolar macrophage cells. A simple approach to obtain alveolar macrophage gene expression is to take the unweighted average across the annotated alveolar macrophage cells in the original study. However, single-cell data are high-dimensional, and annotations for different cells have varying degrees of uncertainty. To better characterize the alveolar macrophage gene expression levels, we took this uncertainty into account by assigning higher weights to cells that we are more certain of them being alveolar macrophages. We quantified such uncertainty with probabilities of cells being alveolar macrophages calculated by HierXGB method [9].

HierXGB is a supervised machine learning algorithm that aims to classify each single cell in the query dataset into one of cell types from a reference dataset. With a pre-defined cell-type hierarchical tree structure, the algorithm annotates the cell from ancestor to one of descendant subtypes iteratively until reaching the bottom layer. For dataset with a clear cell type hierarchy, including Lungmap, it outperforms other state-of-arts methods in terms of both accuracy and efficiency [9]. We performed a comparative analysis using the same IPF dataset in with singleR [1], a widely used method for scRNA-seq data annotation to demonstrate the usefulness of HierXGB method in our setting.

In the analysis, we first had Lungmap and IPF scRNA-seq datasets aligned using batch effect correction [40]. Then the HierXGB model was trained by Lungmap and produced the predictive probability of a cell belonging to alveolar macrophage for the IPF scRNA-seq data. The obtained probabilities were used as the weights when we combined gene expression across cells to obtain cell-type-specific expression summary for each gene per individual.

Generating predictor X_2 based on alveolar macrophage probabilities and outcome Y as the weighted expression average of alveolar macrophage per patient

In our example, we averaged the probability of being in alveolar macrophages across cells within each of the 22 donors and generated a length-22 vector as the predictor $\boldsymbol{X}_2$. A higher $\boldsymbol{X}_2$ indicates that the cells from this donor are more likely from alveolar macrophages. We also used this alveolar macrophage probability to generate outcome Y. For each gene, Y was also a length-22 vector, where the value was the weighted average of counts in this gene across cells. The weights were alveolar macrophage probabilities, and such weighted averages are called pseudo-bulk counts in single cell data analysis. Weighted averages are more robust to different cell numbers per donor. A higher value of Y indicates that this gene expresses highly in alveolar macrophage cells.

Acquiring priors of the contrast between IPF and healthy and other covariates (β) and their variance–covariance matrix Σ,

We use non-informative priors such as $(0, 0, 0)$ and $diag(100, 100, 100)$ for μ_β and Σ_β, respectively, when we have little information on parameters. However, for RNA-sequencing data, bulk RNA-seq data which are characterized by its affordability and wide availability, can serve as good informative priors. Although bulk data may not have the same high-resolution as single-cell data, they still provide the overall expression level of each gene within the targeted tissue. In our analysis, the key parameter is β_1 for the disease group, and we incorporated bulk RNA-seq into its estimation. For each gene, we used the difference in mean expression levels between IPF and control samples as the prior of β_1. We obtained the difference from the coefficient of IPF indicator in a linear regression model adjusting for other covariates including age, smoke, sex, and race. We observed that the sample variance of the difference was relatively small compared with the magnitude of the difference. Directly using it as the prior for the variance of β_1 would result in a very strong prior distribution concentrated around the mean. To offer a prior with more moderate dispersion, we used the squared root of sample variance of difference as the variance for the prior· For intercept and average negative log expression, we had no prior information and hence kept the non-informative priors for (β_0, β_2).

Pathway analysis based on differentially expressed genes from frequentist, Bayesian and hybrid methods

Pathway analysis is usually performed using a set of selected features, in this case, differentially expressed genes [23]. The goal of the analysis is to identify common biological pathways or networks and analyze how they interact to form biological processes. The analysis typically involves comparing a list of genes of interest to a reference database that contains information about functional categories such as biological pathways, molecular interactions, canonical signaling, disease biomarker and other areas of biomedical knowledge. The analysis determines whether the genes in the list are significantly enriched or depleted in any of the categories, compared with what would be observed by chance. A pathway with significantly enriched genes would yield a significantly small p-value. We further studied pathways based on a gene subset that satisfies the following criteria. For a set of differentially expressed genes detected by each method, cut-off points were set to obtain a gene subset with FDR less than 0.01 and absolute estimated difference greater than 0.585. The reference database we used was the R package metabaser that collects all system biology products including MetaCore, MetaDrug, and others [6, 30]. The final pathways were identified based on a p-value less than 0.05. Please see the top10 and detailed summary of pathways in Table 1 and 2 and supplement material S5 and S6.

Results

Comparison of differential expression methods using semi-synthetic dataset

For each method, we obtained p-values and fold changes (FC) in mean expression between two groups ("disease" vs. normal chow) for the genes in each semi-synthetic dataset. Initially, we adjusted the p-values of tested genes for multiple comparisons using the Benjamini–Hochberg procedure [3]. The FC also serves as a crucial metric in determining if a gene is differentially expressed. We applied a 0.05 False Discovery Rate (FDR) cutoff and a 1.5 FC cutoff in this setting. Consequently, a gene was classified as differentially expressed when its adjusted p-value was less than 0.05 and its FC exceeded 1.5.

Power and FDR calculations, as discussed in section "Single-cell RNA sequencing methods comparison using semi-synthetic dataset", are summarized in Fig. 1. Our findings reveal that our proposed hybrid methods

Table 1 Top 10 pathways detected by Bayesian, informative method, ranked by qvalue

Bayesian, informative							
Pathways	**r**	**R**	**n**	**N**	**Zscore**	**pvalue**	**qvalue**
Role of TGF-beta 1 in fibrosis development after myocardial infarction	10	219	38	12,814	11.72026	5.39E−10	8.23E−07
IL-1 beta- and Endothelin-1-induced fibroblast/ myofibroblast migration and extracellular matrix production in asthmatic airways	8	219	40	12,814	8.93905	3.08E−07	0.000173
Cell adhesion_ECM remodeling	9	219	55	12,814	8.403013	3.39E−07	0.000173
TGF-beta-induced fibroblast/ myofibroblast migration and extracellular matrix production in asthmatic airways	9	219	60	12,814	7.961536	7.33E−07	0.00028
Th2 cytokine- and TNF-alpha-induced profibrotic response in asthmatic airway fibroblasts/ myofibroblasts	8	219	52	12,814	7.623879	2.52E−06	0.000771
Immune response_CCL2 signaling	8	219	54	12,814	7.445987	3.39E−06	0.000863
TGF-beta 1-mediated induction of EMT in normal and asthmatic airway epithelium	7	219	44	12,814	7.279623	8.66E−06	0.00189
Development_Inhibition of angiogenesis and regulation of endothelial cell function by PEDF	8	219	64	12,814	6.677022	1.25E−05	0.002377
Immune response_IL-4-responsive genes in type 2 immunity	8	219	70	12,814	6.291138	2.43E−05	0.003892
Role of fibroblasts in the sensitization phase of allergic contact dermatitis	5	219	22	12,814	7.61249	2.89E−05	0.003892

* Threshold: qvalue < 0.05; r: intersection of ontology term with experiment list; R: size of experiment list; n: size of ontology term; N: size of background list; zscore: z-score of enrichment; pvalue: hypergeometric test enrichment p-value; qvalue: FDR-adjusted pvalue

Table 2 Top 10 pathways detected by Hybrid, informative method, ranked by qvalue

Hybrid, informative							
Pathways	**r**	**R**	**n**	**N**	**Zscore**	**pvalue**	**qvalue**
Role of TGF-beta 1 in fibrosis development after myocardial infarction	10	236	38	12,814	11.23695	1.12E−09	1.71E−06
IL-1 beta- and Endothelin-1-induced fibroblast/ myofibroblast migration and extracellular matrix production in asthmatic airways	8	236	40	12,814	8.554393	5.44E−07	0.000324
Cell adhesion_ECM remodeling	9	236	55	12,814	8.026844	6.37E−07	0.000324
TGF-beta-induced fibroblast/ myofibroblast migration and extracellular matrix production in asthmatic airways	9	236	60	12,814	7.598001	1.37E−06	0.000522
Development_Inhibition of angiogenesis and regulation of endothelial cell function by PEDF	9	236	64	12,814	7.289229	2.39E−06	0.000729
Th2 cytokine- and TNF-alpha-induced profibrotic response in asthmatic airway fibroblasts/ myofibroblasts	8	236	52	12,814	7.277826	4.4E−06	0.000978
Immune response_Alternative complement pathway	8	236	53	12,814	7.190269	5.1E−06	0.000978
Immune response_IL-4-responsive genes in type 2 immunity	9	236	70	12,814	6.872979	5.12E−06	0.000978
Immune response_CCL2 signaling	8	236	54	12,814	7.104981	5.89E−06	0.001
TGF-beta 1-mediated induction of EMT in normal and asthmatic airway epithelium	7	236	44	12,814	6.951711	1.41E−05	0.002153

* Threshold: qvalue < 0.05; r: intersection of ontology term with experiment list; R: size of experiment list; n: size of ontology term; N: size of background list; zscore: z-score of enrichment; pvalue: hypergeometric test enrichment p-value; qvalue: FDR-adjusted *p* value

could be optimal when both power and FDR were considered especially with an informative prior. MAST demonstrated high power akin to the hybrid approach, albeit with an inflated FDR exceeding 50%, aligning with Squair et al.'s findings. In contrast, NEBULA exhibited insufficient power and inflated FDR. Finally, none of the pseudo-bulk methods achieved more than 1% power, although the FDR closely approached the nominal level of 5%.

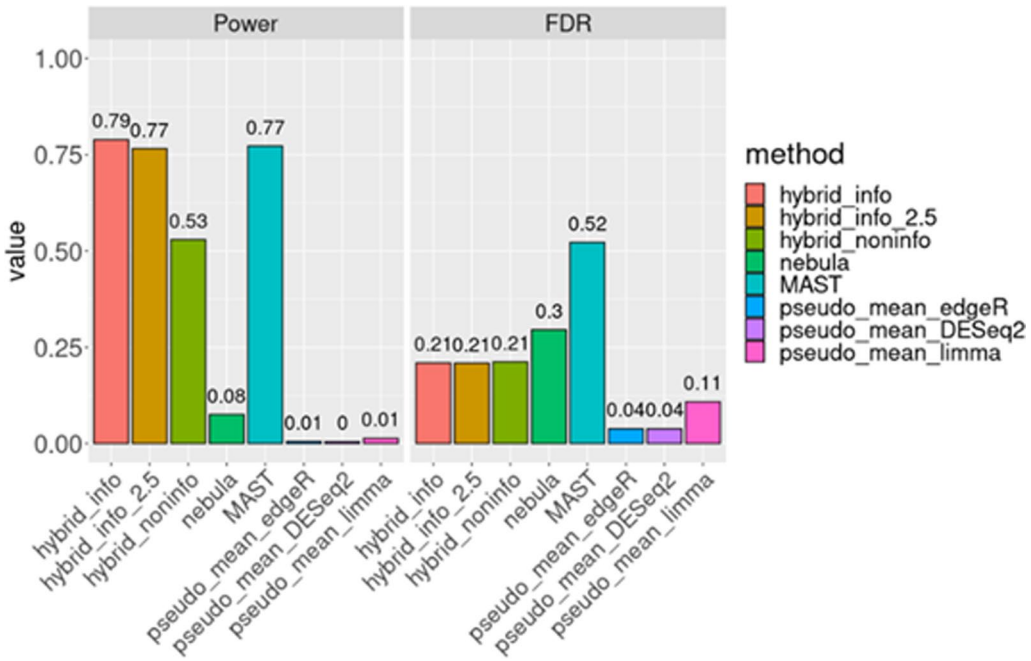

Fig. 1 The comparison of different DE methods using 100 runs of semi-synthetic data. The proposed hybrid methods were implemented with non-informative and different informative priors

Analysis of transforming growth factor *beta* 1 (TGF-β1) gene from frequentist, Bayesian and BFH methods

Through our comparison analysis using singleR and HierXGB methods, the accuracy and mean F1 of alveolar macrophages are 0.897, 0.921 and 0.829, 0.914 for HierXGB and SingleR respectively when compared with the annotation from the original paper. This demonstrated our HierXGB method was at least in par with other popular single cell annotation methods.

Next, we exemplified and compared the frequentist, Bayesian, and hybrid inferences with and without informative prior. TGF-β pathway is well-known in terms of its role in pulmonary fibrosis, therefore, we used the results of TGF-β1 as an example [15]. In the analysis, the outcome or response variable (y) was the weighted average gene expression level per individual (see sections "Acquiring weights for alveolar macrophage cells using HierXGB method" and "Generating predictor based on alveolar macrophage probabilities and outcome Y as the weighted expression average of alveolar macrophage per patient" for details). The two independent variables include 1) whether the individual was in the control or IPF group (X_1) and 2) the average negative log probability of being the macrophage cell (X_2). The model parameters ($\beta_0, \beta_1, \beta_2$) were intercept, coefficients for X_1 and X_2, in the regression model, respectively.

Figure 2 shows boxplots of average of gene expression per person grouped by IPF or control. In this sample 12 patients were in the IPF group and 10 were in the control group. Panel (a) has the average gene expressions of all cells for each person, weighted by the probability of each cell being alveolar macrophage cell. Panel (b) has the average gene expression from the cells that were predicted to be alveolar macrophages based on HierXGB prediction. In both (a) and (b), the expression of TGF-β1 in IPF is lower than control, but not statistically significant. The averages of gene expression for control in (a) and (b) are 1.21 and 1.22, respectively. The averages of gene expression for IPF in (a) and (b) are 0.93 and 0.86, respectively. Numerically the expressions from IPF are lower than from control, but Wilcoxon rank sum test p-values are 0.448 for (a) and 0.419 for (b), both are not statistically significant.

Table 3 is a summary of analysis result for gene TGF-β1 and model coefficient for disease group (i.e. IPF and control) (β_1), including in the columns the coefficient estimate, standard error, 95% confidence interval, and p-value for testing if the estimated coefficient is different from 0. The linear regression also included intercept and gene probability of being microphage data as a covariate. The five rows in Table 3 correspond to 5 inferences about β_1.

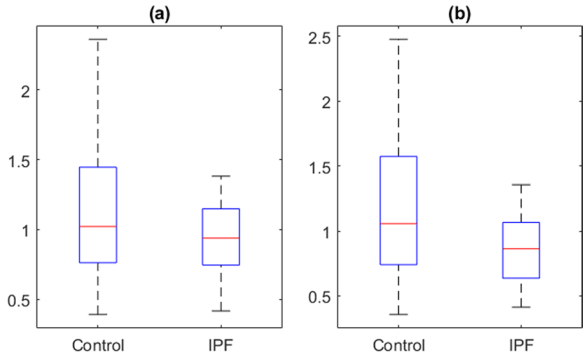

Fig. 2 Boxplot of individual level TGF-β1 gene expression by phenotype in the whole sample (**a**) and predicted alveolar macrophages (**b**)

Frequentist: All model parameters were frequentist parameters, and ordinary least square estimates were reported.

Bayesian inference with non-informative prior: A non-informative prior distribution was imposed on all the parameters $\beta_0, \beta_1, \beta_2$. Mean and standard deviation of the posterior distribution (μ_β^{new}, Σ_β^{new}) were reported.

Hybrid inference with non-informative prior: The same non-informative prior was imposed on β_1, while β_0, β_2 were frequentist parameters.

The Bayesian and hybrid Bayesian inference with informative prior had the Normal distribution with mean −0.31 and variance 0.096 as the prior for β_1, while β_0, β_2 still had the non-informative priors.

The frequentist and Bayesian analyses for the samples with averaged expression across all predicted alveolar macrophages cells by HierXGB had parameters β_0, β_1 only, and the Bayesian inference was based on the same non-informative prior for β_0, β_1 as in Bayesian inference with non-informative prior.

With non-informative prior, the frequentist with Bayesian inferences resulted in similar estimates of β_1. Bayesian inference had slightly wider 95% confidence interval (CI), (−0.692, 0.147), compared with the 95% CI (−0.695, 0.145) from the frequentist inference. The hybrid inference had a similar estimate of β_1 but less standard error (0.144) and shorter 95% CI (−0.557, 0.008) than the Bayesian inference. The hybrid inference with non-informative was marginally significant with p-value 0.057. Given the informative prior, both Bayesian and hybrid inference showed significant effect on β_1, with p-values of 0.017 and 0.005, respectively. The estimates of β_1 were identical (−0.299), but the standard error from hybrid inference (0.106) was less than that from Bayesian inference (0.126), leading to a smaller, more significant p-value from the hybrid inference. This is consistent with

Table 3 The estimation, standard error, 95% confidence interval (95% CI), p-value of the difference between IPF and healthy (β_1) for gene TGF-β from 7 models: frequentist, Bayesian inference with non-informative and informative priors, hybrid inference with non-informative and informative priors for all cells; and frequentist and Bayesian analysis for the predicted alveolar macrophages

Sample	Method	Estimate	Standard error	95% CI	P-value
All cells weighted by alveolar macrophages predictive probability	Frequentist	−0.275	0.214	(−0.695, 0.145)	0.199
	Bayesian, non-informative	−0.273	0.215	(−0.692, 0.147)	0.201
	Hybrid, non-informative	−0.275	0.144	(−0.557, 0.008)	0.057
	Bayesian, informative	−0.299	0.126	(−0.545, −0.053)	0.017
	Hybrid, informative	−0.299	0.106	(−0.506, −0.092)	0.005
Predicted alveolar macrophages	Frequentist	−0.357	0.228	(−0.804, 0.090)	0.117
	Bayesian, non-informative	−0.356	0.228	(−0.803, 0.091)	0.119

the published literature of TGF-β1's critical role for pulmonary fibrosis [15].

We also conducted analysis on the samples with average expression of cells predicted as alveolar macrophage by HierXGB. In this analysis the probability of alveolar macrophage prediction was no longer used but the expression was averaged across identified alveolar macrophages cell types as described in Sect. "scRNA-seq and bulkRNA-seq Data source" using naïve Bayesian approach. This analysis was consistent with the traditional pseudo bulk analysis, ignoring the predictive probability of cell identity. The frequentist regression analysis p-value is equivalent to that from the two-sample t-test, because the independent variable phenotype is binary, and the F-statistic from regression (or ANOVA) is the square of the t-statistic in the t-test. In this analysis, the Bayesian inference with non-informative prior had similar results as frequentist and both were not significant. Such inference was worse than BFH method with informative prior.

As a result, the analysis of TGF-β1 gene indicates that BFH inference outperforms both frequentist and Bayesian inference. The inclusion of cell type predicted probability for all the cells (regardless how small the probability was), and the informative prior were all valuable for identifying potential significant genes.

Interpretation of differentially expression results from frequentist, Bayesian and BFH methods

We applied each of the five methods to the IPF single cell dataset. The hybrid method with an informative prior detected the largest quantity of genes and biological meaningful pathways. Figure 3 summarizes the number of genes detected by each method. See detailed estimation, standard error, and p-value etc. in Table S2-S4. The hybrid method with an informative prior has the highest power with 436 genes detected. Compared with the Bayesian method with an informative prior that discovered 416 genes, the hybrid method detected all of them with an additional 20 genes (Table S1). Among the 20 genes, TREM1 and CCL24 are the most interesting discoveries. Multiple studies have shown their association with IPF. TREM-1 is a receptor expressed on myeloid cells that could serve as an inflammatory biomarker. For example, Dong et al. [11] studied a highly selective inhibitor to suppress TREM-1 expression and inflammation in murine macrophage. A previous study by Xiong et al. [43] found that TREM-1 was upregulated in bleomycin (BLM)-induced pulmonary fibrosis (PF) mouse model. They further discovered a pro-fibrotic effect of TREM-1 in PF, a potential strategy for treating fibrotic diseases could be provided. CCL24 protein promotes immune cell trafficking and activation as well as activities that lead to fibrosis. Kohan et al. [24] revealed that eotaxin-2, the protein encodes by CCL24, stimulated human lung fibroblast proliferation. Mor et al. [28] concluded that CCL24 plays an important role in skin and lung inflammation and fibrosis pathological progression.

The hybrid method with informative discovered most pathways with a total of 38, whereas the Bayesian method discovered 36 (Fig. 4). The top pathway from each method involves TGF-β1 in fibrosis development. See Table 1, 2, 3, 4. TGF-β is a multifunctional cytokine that belongs to the transforming growth factor superfamily and has multiple isoforms, TGF-β1 is one of them. It has been well established that TGF-β1 plays a role in acute respiratory distress syndrome and pulmonary fibrosis [15]. Past publications have studied the role of TGF-β in alveolar macrophages development. For example, Yu et al. [44] revealed that TGF-β plays an essential role in controlling the origin, development,

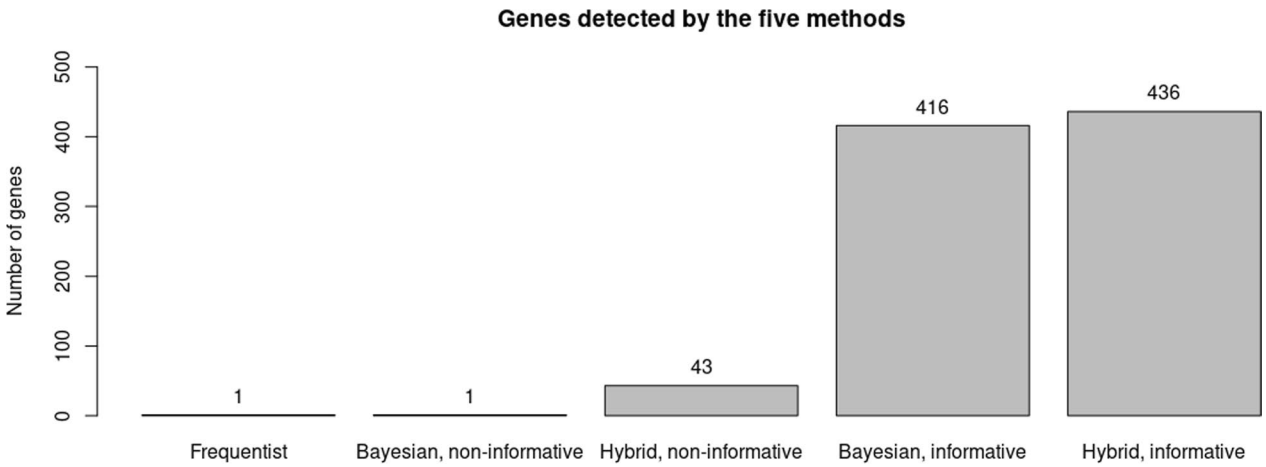

Fig. 3 Genes detected by the five methods (threshold: adjusted *p*-value < 0.01 and absolute value of mean estimation > = 0.585). *P*-value adjust: Benjamini and Yekutieli [4] FDR control

and survival of alveolar macrophages. Woo, Jeong, and Chung [41] reviewed the role of TGF-β in alveolar macrophages development, provided new information and insight into its functions. Grunwell et al. [15] discovered that targeting the TGF-β1 signalling pathway disruption may be a novel therapeutic approach to improve alveolar macrophage function. In comparison, the hybrid method with non-informative method discovered only two pathways, and their linkage to IPF remains unclear (Table 4). In both gene and pathway discoveries, the hybrid method with an informative prior showed supreme detection power against other methods.

Conclusion and discussion

Analysing scRNA-seq data has been a challenging topic, especially given the high cost of running the experiments, which typically results in limited sample sizes. Pseudo-bulk methods, which pool scRNA-seq counts per patient per cell-type, have been commonly used for DE gene detection. However, its performance also relies on the sample size, and hence may lack detection power when sample size is limited. A natural way to overcome this challenge is to borrow information from other studies. Here, we have shown that BFH with informative priors should be considered and has advantages over other approaches. This could be seen in our method comparison using semi-synthetic dataset. BFH can be viewed

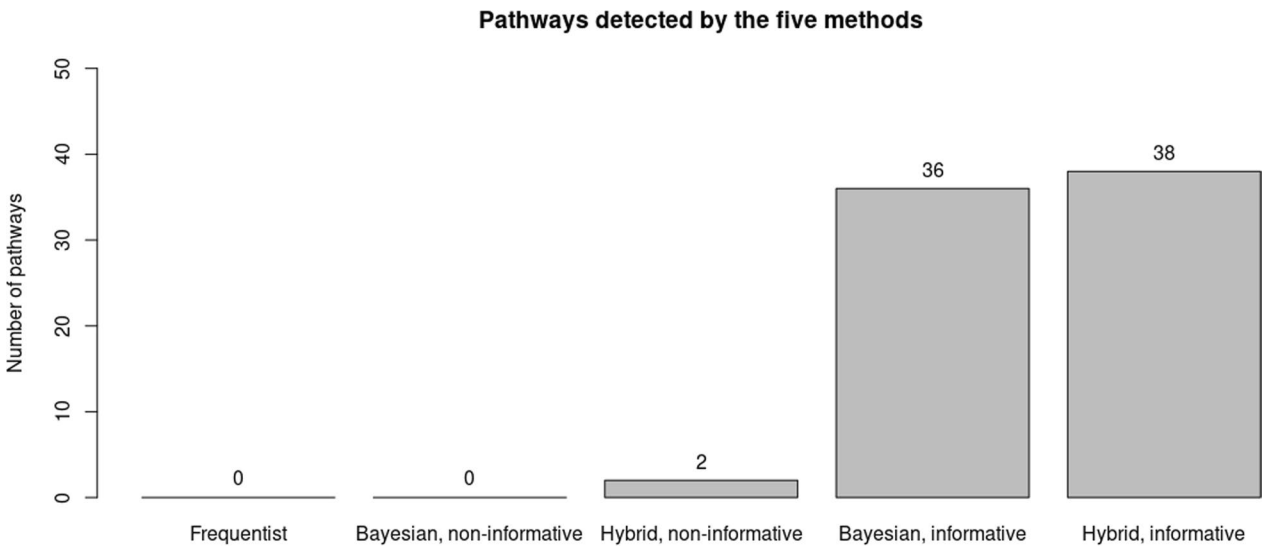

Fig. 4 pathways detected by the five methods based on the genes detected and the threshold for pathway analysis is based on q value < 0.05

Table 4 A detailed list of pathways detected by Hybrid, non-informative method

Hybrid, non-informative							
Pathways	**r**	**R**	**n**	**N**	**Zscore**	**pvalue**	**qvalue**
Putative pathways of activation of classical complement system in major depressive disorder	4	37	28	12,814	13.81785	1.15E−06	0.00175
Development_Role of proteases in hematopoietic stem cell mobilization	3	37	18	12,814	12.95844	1.76E−05	0.013401

* Threshold: qvalue < 0.05; r: intersection of ontology term with experiment list; R: size of experiment list; n: size of ontology term; N: size of background list; zscore: z-score of enrichment; pvalue: hypergeometric test enrichment p-value; qvalue: FDR-adjusted pvalue

as an adaption of Bayesian method and can incorporate prior information with potentially less uncertainty compared with Bayesian methods. As can be seen from the comparison, acquiring valuable priors is crucial for successfully identifying DE genes. Although we only showed one IPF dataset with bulk RNA-seq as prior, which may have inflated FDR, similar to the Bayesian framework, our BFH method can be implemented iteratively, especially when multiple datasets are accessible. In this iterative process, the posterior obtained from previous analyses becomes a more informative prior for the subsequent analysis. Over iterations, the prior and posterior regarding the identification of differentially expressed genes gradually converge. This convergence significantly improves our ability to pinpoint the correct genes associated with the disease using single-cell RNA sequencing (scRNA-seq) data with potential reduction in FDR.

In our BFH analysis of the IPF study, we used pseudo-bulk summarization as the response variable. However, the way to calculate pseudo-bulk is still a topic of discussion in the field. While many researchers have used the annotated cell types directly and summarized the data within a particular annotated cell type, such summarization may potentially lose information since the annotation is based on classifiers with certain thresholds to define the cell types. In our study, we derived the cell-type-specific probability for each cell instead of relying on a classifier to define the cell types. We used this probability as the weight to summarize the data to avoid loss of useful information.

To boost the power of our coefficient estimate, we used bulk RNA-seq data as the prior for each gene. Bulk RNA-seq data are not cell-specific, thus if the cell of interest is relatively scarce, they may not be able to provide useful information. Literature suggested that alveolar macrophages were abundant in the lung and could play a pivotal role in immunity [38]. Although bulk RNA-seq data may not be as informative as scRNA-seq dataset to be used as prior, they offer several advantages. First, bulk data have better coverage than scRNA-seq data, thus providing prior information on a more compressive gene list than a typical scRNA-seq experiment. Second, it is relatively inexpensive and readily available. As we have summarized the scRNA-seq data into pseudo-bulk format, the prior derived from bulk RNA-seq data is compatible with our scRNA-seq data. In addition, we transformed the sample variance for β_1 so that the prior would have better dispersion. Properly setting priors remains as an interesting topic and should be explored further. The BFH method inherits flexibility from the Bayesian framework and can be used iteratively to integrate the current results as new prior information with the new data when appropriate.

Our case example demonstrated the substantial increase in detection power of BFH framework when using informative priors. When non-informative priors were employed, either no differentially expressed genes were identified or only a small number were found. The use of informative priors significantly increased the detection power, as evidenced by the reasonable pathway identified for IPF in terms of the underlying mechanism. The work reinforces the importance for TGF-β pathway and cytokines such as TNF and IL1/IL4, which are well-known for their roles in the IPF mechanism [5, 14, 31]. Consequently, our work brings valuable biological insight into the IPF disease for researchers.

A potential limitation of the BFH method is its heavy reliance on informative priors. In situations where relevant bulk RNA-seq or scRNA-seq data are unavailable, alternative data types such as methylation data could be considered as prior information. However, developing such priors needs biological justification and consideration of how to align such data types to pseudo-bulk format of scRNA-seq data. When alternative data types are unavailable, the use of non-informative prior for parameters is inevitable. As discussed in literature [18, 19], using Bayesian analysis with non-informative prior can lead to estimation bias and incorrect p-values if the sample size is relatively small.

Despite its limitations, the BFH method is a flexible approach with the capability to incorporate informative prior to enhance detection power. The current framework of the BFH method is implemented using conjugate priors, which reduces the computation time and makes it a suitable method for high throughput analyses in the future.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-024-00638-0.

Additional file 1: Table S1: The estimation, standard error, 95% confidence interval (95% CI), p-value, adjusted p-value of β1 from the gene detection difference between Hybrid and Bayesian method with informative priors; Table S2: The estimation, standard error, 95% confidence interval (95% CI), p-value, adjusted p-value of β1 from the detailed list of genes detected by Hybrid, non-informative method (43 genes); Table S3: The estimation, standard error, 95% confidence interval (95% CI), p-value, adjusted p-value of β1 from the detailed list of genes detected by Bayesian, informative method (416 genes); Table S4: The estimation, standard error, 95% confidence interval (95% CI), p-value, adjusted p-value of β1 from the detailed list of genes detected by Hybrid, informative method (436 genes); Table S5: A detailed list of pathways detected by Bayesian, informative method (36 pathways); Table S6: A detailed list of pathways detected by Hybrid, informative method (38 pathways).

Acknowledgements
We would like to thank participants and researchers to make lungmap dataset, IPF scRNA-seq dataset, and IPF bulk RNA-seq dataset publicly available.

Author contributions
Conceptualization, G.H., and Y.L.; methodology, G.H., J.F, D.Y, and Y.L; formal analysis, G.H., Z.S, J.F., D.Y., X.C., and Y.L; data curation, G.H., and Z.S.; writing—original draft preparation, G.H., Z.S, J.F., D.Y., X.C., L.W., and Y.L; writing—review and editing, G.H., Z.S, J.F., D.Y., X.C., L.W, and Y.L; All authors have read and agreed to the published version of the manuscript.

Funding
This research is partially funded by DHHS-NIH-National Institute of Environmental Health Sciences, grant P30ES029067.

Data Availability
Lungmap dataset could be downloaded from GEO database as GSE161382. IPF scRNA-seq dataset could be downloaded from GEO database as GSE135893. IPF bulk RNA-seq dataset could be downloaded from GEO database as GSE150910. All the relevant code could be downloaded at https://github.com/hangangtrue/HB_singlecell.

Declarations

Competing interests
Z.S, J.F., D.Y., X.C., and Y.L are employees and stockholders of Eli Lilly and Company.

Received: 25 September 2023 Accepted: 12 June 2024
Published online: 20 June 2024

References

1. Aran D, Looney AP, Liu L, Esther Wu, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72.
2. Bargagli E, Prasse A, Olivieri C, Muller-Quernheim J, Rottoli P. Macrophage-derived biomarkers of idiopathic pulmonary fibrosis. Pulmon Med. 2011.
3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57:289–300.
4. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat 2001;1165–88.
5. Borthwick LA. The IL-1 cytokine family and its role in inflammation and fibrosis in the lung. In: Seminars in immunopathology. Springer; 2016. pp. 517–34.
6. Bureeva S, Zvereva S, Romanov V, Serebryiskaya T. 2009. Manual annotation of protein interactions. In: Protein networks and pathway analysis 2016;75–95.
7. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
8. Cao Y, Yang P, Yang JYH. A benchmark study of simulation methods for single-cell RNA sequencing data. Nat Commun. 2021;12:6911.
9. Chang X, Sun Z, Yan D, Wang W, Liu Y. HierXGB—hierarchical classification of single cells by XGBoost and KNN. 2023; Manuscript in preparation.
10. Das S, Rai A, Merchant ML, Cave MC, Rai SN. A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. Genes. 2021;12:1947.
11. Dong L, Zhou Y, Zhu Z-Q, Liu T, Duan J-X, Zhang J, Li P, Hammcok BD, Guan C-X. Soluble epoxide hydrolase inhibitor suppresses the expression of triggering receptor expressed on myeloid cells-1 by inhibiting NF-kB activation in murine macrophage. Inflammation. 2017;40:13–20.
12. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, Juliana McElrath M, Prlic M. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:1–13.
13. Furusawa H, Cardwell JH, Okamoto T, Walts AD, Konigsberg IR, Kurche JS, Bang TJ, Schwarz MI, Brown KK, Kropski JA. Chronic hypersensitivity pneumonitis, an interstitial lung disease with distinct molecular signatures. Am J Respir Crit Care Med. 2020;202:1430–44.
14. Groves AM, Johnston CJ, Misra RS, Williams JP, Finkelstein JN. Effects of IL-4 on pulmonary fibrosis and the accumulation and phenotype of macrophage subpopulations following thoracic irradiation. Int J Radiat Biol. 2016;92:754–65.
15. Grunwell JR, Yeligar SM, Stephenson S, Ping XD, Gauthier TW, Fitzpatrick AM, Lou Ann S, Brown. TGF-β1 suppresses the type I IFN response and induces mitochondrial dysfunction in alveolar macrophages. J Immunol. 2018;200:2115–28.
16. Gupta K, Lalit M, Biswas A, Sanada CD, Greene C, Hukari K, Maulik U, Bandyopadhyay S, Ramalingam N, Ahuja G. Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-seq data. Genome Res. 2021;31:689–97.
17. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung M-I, Taylor CJ, Jetter C. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv 2020;6: eaba1972.
18. Han G, Huang Y, Yuan Ao. Bayesian-frequentist hybrid approach for skew-normal nonlinear mixed-effects joint models in the presence of covariates measured with errors. Stat Interface. 2018;11:223–36.
19. Han G, Santner TJ, Lin H, Yuan Ao. Bayesian-frequentist hybrid inference in applications with small sample sizes. Am Stat. 2023;77:143–50.
20. He L, Davila-Velderrain J, Sumida TS, Hafler DA, Kellis M, Kulminski AM. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. Commun Biol. 2021;4:629.
21. Keren-Shaul H, Spinrad A, Weiner A, Matcovitch-Natan O, Dvir-Szternfeld R, Ulland TK, David E, Baruch K, Lara-Astaiso D, Toth B. A unique microglia type associated with restricting development of Alzheimer's disease. Cell. 2017;169(1276–90): e17.
22. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.
23. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8: e1002375.
24. Kohan M, Puxeddu I, Reich R, Levi-Schaffer F, Berkman N. Eotaxin-2/CCL24 and eotaxin-3/CCL26 exert differential profibrogenic effects on human lung fibroblasts. Ann Allergy Asthma Immunol. 2010;104:66–72.
25. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:1–17.
26. Li Y, Ge X, Peng F, Li W, Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 2022;23:79.

27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
28. Mor A, Salto MS, Katav A, Barashi N, Edelshtein V, Manetti M, Levi Y, George J, Matucci-Cerinic M. Blockade of CCL24 with a monoclonal antibody ameliorates experimental dermal and pulmonary fibrosis. Ann Rheum Dis. 2019;78:1260–8.
29. Murphy AE, Skene NG. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat Commun. 2022;13:7851.
30. Nikolsky Y, Kirillov E, Zuev R, Rakhmatulin E, Nikolskaya T. 'Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform'. In: Protein networks and pathway analysis. 2009;177–96.
31. Redente EF, Keith RC, Janssen W, Henson PM, Ortiz LA, Downey GP, Bratton DL, Riches DWH. Tumor necrosis factor-α accelerates the resolution of established pulmonary fibrosis in mice by targeting profibrotic lung macrophages. Am J Respir Cell Mol Biol. 2014;50:825–37.
32. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9:284.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
34. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, Hudelle R, Qaiser T, Matson KJE, Barraud Q. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12:5692.
35. Steuernagel L, Lam BYH, Klemm P, Dowsett GKC, Bauder CA, Tadross JA, Hitschfeld TS, del Rio A, Martin WC, De Solis AJ. HypoMap—a unified single-cell gene expression atlas of the murine hypothalamus. Nat Metab. 2022;4:1402–19.
36. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015;25:1491–8.
37. den Berge V, Koen FP, Soneson C, Love MI, Risso D, Vert J-P, Robinson MD, Dudoit S, Clement L. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19:1–17.
38. van Eeden, Stephan F, Sin DD. Lung Macrophages: Pivotal Immune Effector Cells Orchestrating Acute and Chronic Lung Diseases. In: Macrophages-Celebrating 140 Years of Discovery. 2022; (IntechOpen).
39. Wang A, Chiou J, Poirion OB, Buchanan J, Valdez MJ, Verheyden JM, Hou X, Kudtarkar P, Narendra S, Newsome JM. Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. Elife. 2020;9: e62522.
40. Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E. Integrative inference of brain cell similarities and differences from single-cell genomics. BioRxiv 2018;459891.
41. Woo YD, Jeong D, Chung DH. Development and functions of alveolar macrophages. Mol Cells. 2021;44:292.
42. Wu H, Villalobos RG, Yao X, Reilly D, Chen T, Rankin M, Myshkin E, Breyer MD, Humphreys BD. Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. Cell Metab. 2022;34(1064–78): e6.
43. Xiong J-B, Duan J-X, Jiang N, Zhang C-Y, Zhong W-J, Yang J-T, Liu Y-B, Feng Su, Zhou Y, Li D. TREM-1 exacerbates bleomycin-induced pulmonary fibrosis by aggravating alveolar epithelial cell senescence in mice. Int Immunopharmacol. 2022;113: 109339.
44. Yu X, Buttgereit A, Lelios I, Utz SG, Cansever D, Becher B, Greter M. The cytokine TGF-β promotes the development and homeostasis of alveolar macrophages. Immunity. 2017;47(903–12): e4.
45. Yuan A. Bayesian frequentist hybrid inference; 2009.
46. Zhang L, Wang Yi, Guorao Wu, Xiong W, Weikuan Gu, Wang C-Y. Macrophages: friend or foe in idiopathic pulmonary fibrosis? Respir Res. 2018;19:1–10.
47. Zhang M, Liu Si, Miao Z, Han F, Gottardo R, Sun W. IDEAS: individual level differential expression analysis for single-cell RNA-seq data. Genome Biol. 2022;23:1–17.
48. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
49. Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun. 2021;12:738.

Publisher's Note