Genomics Proteomics Bioinformatics 19 (2021) 330–341

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

scLM: Automatic Detection of Consensus Gene Clusters Across Multiple Single-cell Datasets

Qianqian Song [1,2], Jing Su [1,3], Lance D. Miller [1,2], Wei Zhang [1,2,*]

[1] *Center for Cancer Genomics and Precision Oncology, Wake Forest Baptist Comprehensive Cancer Center, Wake Forest Baptist Medical Center, Winston Salem, NC 27157, USA*
[2] *Department of Cancer Biology, Wake Forest School of Medicine, Winston Salem, NC 27157, USA*
[3] *Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Received 19 December 2019; revised 11 August 2020; accepted 27 October 2020
Available online 24 December 2020

Handled by Wei Lin

KEYWORDS

Single-cell RNA sequencing;
Consensus clustering;
Latent space;
Markov Chain Monte Carlo;
Maximum likelihood approach

Abstract In gene expression profiling studies, including **single-cell RNA sequencing** (scRNA-seq) analyses, the identification and characterization of co-expressed genes provides critical information on cell identity and function. Gene co-expression clustering in scRNA-seq data presents certain challenges. We show that commonly used methods for single-cell data are not capable of identifying co-expressed genes accurately, and produce results that substantially limit biological expectations of co-expressed genes. Herein, we present single-cell Latent-variable Model (scLM), a gene co-clustering algorithm tailored to single-cell data that performs well at detecting gene clusters with significant biologic context. Importantly, scLM can simultaneously cluster multiple single-cell datasets, *i.e.*, **consensus clustering**, enabling users to leverage single-cell data from multiple sources for novel comparative analysis. scLM takes raw count data as input and preserves biological variation without being influenced by batch effects from multiple datasets. Results from both simulation data and experimental data demonstrate that scLM outperforms the existing methods with considerably improved accuracy. To illustrate the biological insights of scLM, we apply it to our in-house and public experimental scRNA-seq datasets. scLM identifies novel functional gene modules and refines cell states, which facilitates mechanism discovery and understanding of complex biosystems such as cancers. A user-friendly R package with all the key features of the scLM method is available at https://github.com/QSong-github/scLM.

Introduction

Co-expressed genes work in concert in biological pathways and processes [1–3]. Such genes are involved in crucial biological activities including immune cell activation [4,5], cellular epithelial-mesenchymal transition (EMT) [6], and transcription factor-mediated gene regulatory networks and signaling path-

* Corresponding author.
E-mail: wezhang@wakehealth.edu (Zhang W).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.09.002

ways [7,8]. Co-expression of genes based on similarities among their expression profiles, has been a primary way to unravel gene-gene relationships and facilitate functional annotation [9–12]. Therefore, identification of co-expressed genes provides functional insights into underlying cellular and molecular mechanisms in normal and disease processes.

The recently developed single-cell RNA sequencing (scRNA-seq) technology provides high resolution of gene expression at the single-cell level [13], yet presents certain challenges for gene expression analysis [14]. In contrast to bulk RNA-seq, single-cell data have been shown to exhibit a characteristic negative binomial (NB) distribution pattern [15–18], wherein genes suffer from stochastic dropouts and over-dispersion problems. Dropouts, or genes that exhibit excessive zero values [19–21], represent a special type of missing value, which can be caused by low RNA input or stochastic expression fluctuation at the single-cell level. Over-dispersion relates to the substantially large cell-to-cell variability in gene expression profiles which likely arises from technical noise stemming from low input RNA and PCR amplification bias [22].

Rapid advances in scRNA-seq technologies have made it feasible to perform population-scale studies in which the transcriptome is measured for thousands of single cells from multiple samples or conditions [23–27]. This in turn has amplified the need for versatile gene co-expression approaches that not only address the unique challenges of scRNA-seq data, but also the challenges of dataset integration including batch effects, technical variations (*e.g.*, mRNA quality and pre-amplification efficiency), and extrinsic biological variabilities.

Classical methods designed for analysis of bulk transcriptome data such as weighted gene co-expression network analysis (WGCNA) [28] and Clust [29] are not designed to account for the unique characteristics of scRNA-seq data. Some network-based approaches for single-cell data, including Single-Cell rEgulatory Network Inference and Clustering (SCENIC) [30], Cell Specific Network (CSN) [31], and Left Truncated Mixture Gaussian (LTMG) [32], can detect gene co-expression modules as part of the network reconstruction. However, these methods do not account for the technical noise and extrinsic variance among multiple samples. Therefore, there is a clear need to develop a tailored and effective method for scRNA-seq data to extract "consensus" co-expressed genes [11], that is, to extract the genes that are consistently co-expressed in each of the multiple datasets.

Herein, we have developed a novel method, single-cell Latent-variable Model (scLM), to simultaneously extract co-expressed genes that exhibit consensus behaviors from multiple single-cell datasets. The scLM method accounts for both cell-level covariates and sample-level batch effects. We assessed the performance of scLM in both simulated data and experimental data. scLM achieves the best performance over other commonly used methods. We then applied scLM to our in-house scRNA-seq data generated from four non-small cell lung cancer (NSCLC) tumor tissues and their corresponding adjacent normal tissues. The scLM method identified tumor-specific co-expressed gene modules with significant prognostic values. Furthermore, these co-expression modules contributed to the subtle characterization of lung tumor cell states. In addition, we applied scLM to analyze a set of malignant cells from NSCLC, head and neck squamous cell carcinoma (HNSCC), and melanoma. We discovered a common co-expressed gene program across different cancer types, providing insights into fundamental mechanisms of carcinogenesis.

Method

scLM

We proposed a latent-variable model to explicitly disentangle different sources of variabilities in population-scale scRNA-seq data. Our goal was to perform simultaneous detection of co-expressed genes across multiple single-cell conditions/-datasets. Specifically, let x_{ijk} denotes the gene expression level experimentally measured for the i-th gene ($i \in \{1, \cdots, m\}$) in the j-th cell ($j \in \{1, \cdots, n_k\}$) in condition/dataset $k (k \in \{1, \cdots, K\})$.

As multiple recent studies [15–18] showed that the expression of most genes in single-cell data is sufficiently captured by NB distribution, NB model is chosen as an appropriate model to formulate single-cell data. It is supported by the physical modeling of bursting gene expression [18,33] and is also commonly used in scRNA-seq analysis [15–18]. Therefore, without loss of generality, we assumed that the measured gene expression x_{ijk} for cell j in dataset k follows the NB distribution $NB(p, \gamma)$, which has the probability function as:

$$f(X = x_{ijk}, p, \gamma) = \frac{\Gamma(x_{ijk} + \gamma)}{\Gamma(\gamma)\Gamma(1 + x_{ijk})}(1-p)^{\gamma} p^{x_{ijk}} \tag{1}$$

If μ, θ, and σ^2 represent the mean, dispersion, and variance of this NB distribution, then we have

$$\mu = \frac{p\gamma}{1-p},\ \theta = \gamma,\ \sigma^2 = \frac{p\gamma}{(1-p)^2} \tag{2}$$

also,

$$p = \frac{\mu}{\mu + \theta},\ \gamma = \theta \tag{3}$$

Therefore, the probability function converts to:

$$f(x_{ijk}; \theta, u) = \frac{\Gamma(x_{ijk} + \theta)}{\Gamma(\theta)\Gamma(1 + x_{ijk})}\left(\frac{\theta}{\theta + u}\right)^{\theta}\left(\frac{u}{\theta + u}\right)^{x_{ijk}} \tag{4}$$

As u and θ are regarding different genes ($i \in \{1, \cdots, m\}$) and batches ($k \in \{1, \cdots, K\}$), we have

$$f(x_{ijk}; \theta_{ik}, u_{ik}) = \frac{\Gamma(x_{ijk} + \theta_{ik})}{\Gamma(\theta_{ik})\Gamma(1 + x_{ijk})}\left(\frac{\theta_{ik}}{\theta_{ik} + u_{ik}}\right)^{\theta_{ik}}\left(\frac{u_{ik}}{\theta_{ik} + u_{ik}}\right)^{x_{ijk}} \tag{5}$$

Herein, u_{ik} represents the estimation for the intrinsic gene expression level across all cells in sample k, θ_{ik} is the dispersion parameter, and $\sigma_{ik}^2 = u_{ik} + \frac{u_{ik}{}^2}{\theta_{ik}}$ represents the square deviation of the observed gene expression level across cells in this sample.

Let $z_i = (z_{i1}, \cdots, z_{i\lambda})'$ be a vector consisting of λ unobserved latent variables that are shared by K different datasets. We assumed the generalized linear model (GLM) below

$$u_{ik} \sim \alpha_{jk} + \beta_{jk} z_i \tag{6}$$

which was used to distinguish the intrinsic biological signals z_i from the extrinsic variabilities (α_{jk} and β_{jk}) including the technical variances at the cell level (j) and batch effects at the sample level (k). That is, the u_{ik} is composed of the intrinsic biological signals of gene i captured by latent variables z_i regardless of the

confounding variabilities at the cell level and sample level, while variances due to technical biases and batch effects are captured by offsets α_{jk} and scale factors β_{jk}. Since z_i is the same for specific gene i, and u_{ik} is estimated from observed counts x_{ijk}, we further turned the formula into

$$u(x_{ijk}|z_i) \sim \alpha_{jk} + \beta_{jk} z_i \tag{7}$$

To alleviate the impact of extreme values, we utilized logarithm form in the linear model that has been frequently used [34–38] in scRNA-seq data, *i.e.*, the GLM,

$$\log u(x_{ijk}|z_i) = \alpha_{jk} + \beta_{jk} z_i \tag{8}$$

where $u(x_{ijk}|z_i)$ is the conditional mean of x_{ijk} given z_i. In this way, the original gene expression data were projected into a λ-dimensional latent space Z by the GLM, with the technical biases and batch effects removed during the projection. In this latent space, the expression level of gene i is represented as z_i. Since genes sharing similar expression patterns are located close to each other, a group of co-expressed genes will form a cluster in the latent space. Thus, different groups of co-expressed gene modules can be identified through clustering of the latent variables (**Figure 1**).

To estimate the parameters in our model, we used the maximum likelihood approach. As is assumed above that x_{ijk} follows the NB distribution, the conditional log-likelihood function of x_{ijk} can be written as:

$$\log f(x_{ijk}|z_i, \alpha_{jk}, \beta_{jk}, \theta_{ik}) = \frac{\Gamma(\theta_{ik} + x_{ijk})}{\Gamma(\theta_{ik})\Gamma(1 + x_{ijk})} \left(\frac{u_{ik}}{\theta_{ik} + u_{ik}}\right)^{x_{ijk}} \left(\frac{\theta_{ik}}{\theta_{ik} + u_{ik}}\right)^{\theta_{ik}} \tag{9}$$

in which,

$$u_{ik} = \exp(\alpha_{jk} + \beta_{jk} z_i)$$

For the latent variable z_i, $f(z_i)$ represents the density function of the standard multivariate normal distribution N(0, I_λ). Therefore, the joint log-likelihood of (x_{ijk}, z_i) can be written as

$$l(x_{ijk}, z_i; \alpha_{jk}, \beta_{jk}) = \sum_{i=1}^{m} \sum_{j=1}^{n_k} \sum_{k=1}^{K} \{\log f(x_{ijk}|z_i, \alpha_{jk}, \beta_{jk}) + \log f(z_i)\} \tag{10}$$

To control model complexity and overfitting, we applied the least absolute shrinkage and selection operator (LASSO, L1-norm penalty), to the following penalized joint log-likelihood estimation:

$$l(x_{ijk}, z_i; \alpha_{jk}, \beta_{jk}) - \sum_{k=1}^{K} \sum_{j=1}^{n_k} \partial_k ||\beta_{jk}||_1$$

Then the above parameters are estimated by maximizing the penalized joint log-likelihood function, that is, maximizing the following penalized joint log-likelihood,

$$\begin{aligned} &\max_{\alpha_{jk}, \beta_{jk}} l(x_{ijk}, z_i, \alpha_{jk}, \beta_{jk}) - \sum_{k=1}^{K} \sum_{j=1}^{n_k} \partial_k ||\beta_{jk}||_1 \\ &= \max_{\alpha_{jk}, \beta_{jk}} \sum_{i=1}^{m} \sum_{j=1}^{n_k} \sum_{k=1}^{K} \{\log f(x_{ijk}|z_i, \alpha_{jk}, \beta_{jk}) + \log f(z_i)\} \\ &\quad - \sum_{k=1}^{K} \sum_{j=1}^{n_k} \partial_k ||\beta_{jk}||_1 \end{aligned} \tag{11}$$

where the summation is due to the conditional independence assumption of x_{ijk} given z_i.

To estimate the parameters α_{jk} and β_{jk}, we solved the following optimization problem conditional on z_i,

$$\min_{\alpha_{jk}, \beta_{jk}} \sum_{i=1}^{m} \sum_{j=1}^{n_k} \sum_{k=1}^{K} \log f(x_{ijk}|z_i, \alpha_{jk}, \beta_{jk}) + \sum_{k=1}^{K} \sum_{j=1}^{n_k} \partial_k ||\beta_{jk}||_1 \tag{12}$$

here we used the coordinate descent algorithm provided in [39], therefore optimized the above log-likelihood function. Herein

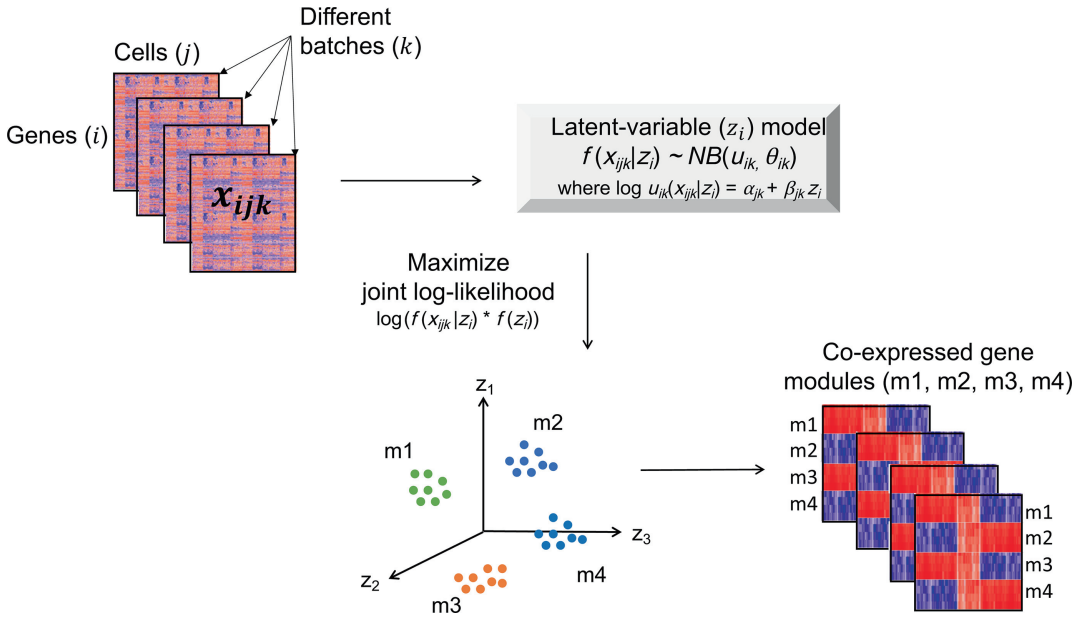

Figure 1 Schematic of scLM for identifying consensus co-expressed gene clusters across multiple datasets

Schematic representation of how consensus co-expressed genes across multiple datasets could be discovered by scLM. The gene expression profiles of individual cells were disentangled by the latent variables representing the intrinsic biological signals, and the related coefficients reflecting the technical variances. scLM, single-cell Latent-variable Model.

the update of parameters α_{jk} and β_{jk} depend on z_i. As the latent variables z_i were not observable in our model, we used the Markov Chain Monte Carlo (MCMC) simulation to iteratively update z_i, for maximizing the penalized joint log-likelihood. That is, we replaced the value in the parameter updates by its expectation with respect to z_i, through repeatedly sampling the latent variables z_i from the following joint posterior distribution, *i.e.*,

$$f(z_i)\prod_{j=1}^{n_k}\prod_{k=1}^{K} f\left(x_{ijk}|z_i,\alpha_{jk},\beta_{jk}\right)$$

With the estimated latent variables z_i, that is, with the genes projected into the latent space, we clustered genes that were projected in the latent space to identify co-expressed genes. Here we used K-means clustering to divide genes into λ clusters based on the latent variables z_i. The parameter λ can be either determined according to the Bayesian information criterion (BIC), or chosen by user's preference.

Data generation in simulation studies

Based on the single-cell data characteristics, we used the NB distribution to simulate two synthetic cohorts (synthetic cohort 1 and synthetic cohort 2). Each synthetic cohort contained 9 sets of simulated gene expression data with an increasing number of datasets (D1–D9). That is, D1 contained one individual dataset ($n = 1$), D2 contained two individual datasets ($n = 2$), ..., and D9 contained nine individual datasets ($n = 9$). Each individual dataset contained 180 genes belonging to three clusters, with 60 co-expressed genes in each of the three clusters. For each gene cluster $c \in \{1, 2, 3\}$ in batch n, their gene expression was sampled from the NB distribution $NB(u_{cn}, \theta_{cn})$, where u_{cn} and θ_{cn} referred to the mean and deviation, respectively. Different gene clusters had different values of u_{cn} and θ_{cn}. Full expression values and cluster membership for these datasets were provided in the scLM example data.

Additionally, we utilized the Splatter package [40] to generate another two synthetic cohorts (synthetic cohort 3 and synthetic cohort 4) of simulated data with dropout effects, which more accurately recapitulated actual scRNA-seq data distributions. Specifically, we adjusted the batch parameters "batch.facLoc" and "batch.facScale" as 1 and generated 16 different batches of data. Each batch consisted of 240 cells, and 240 genes constituting four groups of co-expressed genes as the group truth, which was achieved by adjusting the "de.prob" parameter. We also added the dropout effects in these simulation data by setting "experiment" for global dropout and the "dropout.mid" parameter. These 16 batches of data made up the synthetic cohorts 3 and 4. Full expression values were provided in the scLM example data.

In-house and public single-cell data

In-house dataset

Fresh tumor and adjacent normal tissues from four NSCLC patients were collected by the Tumor Tissue and Pathology Shared Resource (TTPSR) into Miltenyi Tissue Storage Medium (Catalog No. 130-100-008, Miltenyi Biotec, San Diego, CA). Tissues were then processed to single-cell suspensions using the Miltenyi Human Tumor Dissociation Kit (Catalog No. 130-095-929, Miltenyi Biotec) and the gentleMACS Octo Dissociator with Heaters (Catalog No. 130-096-427, Miltenyi Biotec). Red blood cells were removed by negative selection using Miltenyi CD235a (Glycophorin A) microbeads (Catalog No. 130-050-501, Miltenyi Biotec) and LS Columns (Catalog No. 130-042-401, Miltenyi Biotec). Recovered cell numbers were determined by trypan blue exclusion using a LUNA II automated cell counter (Catalog No. L40001, Logos Biosystems, Annandale, VA). In preparation for scRNA-seq, cells were thawed and washed according to the demonstrated protocol developed for human peripheral blood mononuclear cells (PBMCs) by 10X Genomics (San Francisco, CA).

All scRNA-seq procedures were performed by the Cancer Genomics Shared Resource (CGSR) of the Wake Forest Baptist Medical Center Comprehensive Cancer Center (WFBMC-CCC). Viable cells in suspensions were loaded into wells of a 10X Genomics Chromium Single Cell A Chip Kit (Catalog No. PN-120236, 10X Genomics). Single-cell gel beads in emulsion (GEMs) were created on a Chromium Single Cell Controller and scRNA-seq libraries were prepared using the Chromium Single Cell 3′ Library and Gel Bead Kit v2 (Catalog No. PN-120237, 10X Genomics). Sequencing libraries were loaded at 1.3 PM on an Illumina NextSeq500 with a High Output 150 cycle Kit (Catalog No. FC-404-2002, Illumina, San Diego, CA) for paired-end sequencing. A total of 11,813 single cells were captured, with the number of cells recovered per channel ranging from 369 to 2502. Low-quality cells were discarded if the cell with expressed genes was smaller than 200. Only malignant cells from four tumor samples and epithelial cells from three adjacent normal samples were used in this study. The scRNA-seq data were deposited in the Gene Expression Omnibus (GEO) of National Center for Biotechnology Information (NCBI) database (GEO: GSE117570) at https://onlinelibrary.wiley.com/doi/full/10.1002/cam4.2113 [41].

Melanoma dataset

We downloaded the expression matrix data of melanoma from the GEO of NCBI database (GEO: GSE72056) at https://www.ncbi.nlm.nih.gov/pubmed/27124452 [42]. This dataset included expression profiles of 23,689 genes in 4645 cells from 19 melanoma tumors. These cells included both malignant cancer cells and non-malignant cells. For the input matrix of scLM, a sample is excluded if it contains < 200 cells, and a gene is excluded from the input matrix if it is expressed in < 300 cells.

HNSCC dataset

We downloaded the expression matrix data of the HNSCC dataset from the GEO of NCBI database (GEO: GSE103322) at https://www.sciencedirect.com/science/article/pii/S0092867417312709 [6]. This dataset consisted of 5902 cells from 18 patient samples after initial quality controls, including 2215 malignant and 3363 non-malignant cells. For our analyses, we used the samples with more than 200 malignant cells and genes expressed in over 300 cells as the input matrix.

Breast cancer dataset

We downloaded the expression matrix data of breast cancer (BR) scRNA-seq dataset from the GEO of NCBI database (GEO: GSE118390) at https://www.nature.com/articles/

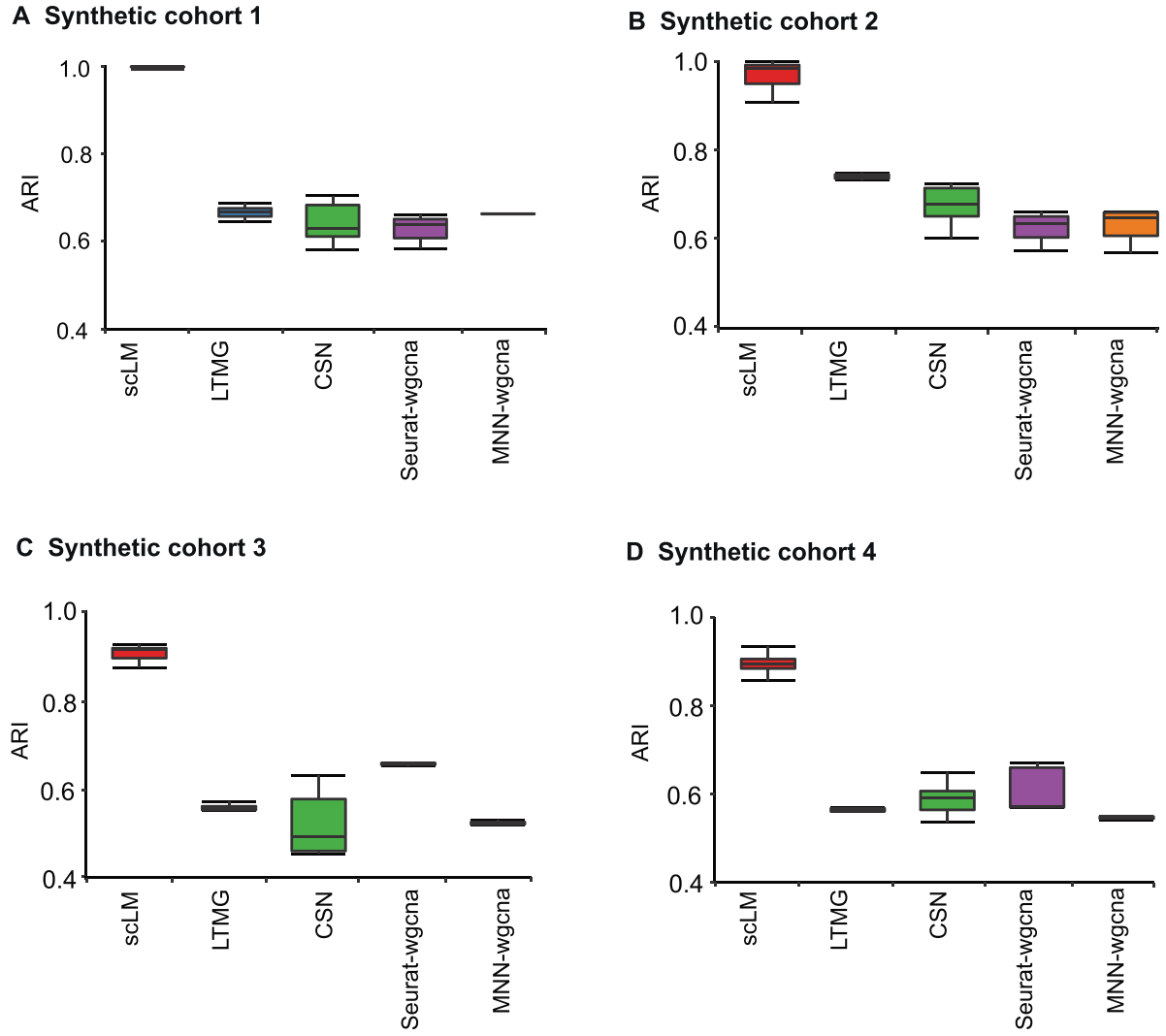

Figure 2 Performance evaluation on simulation data

scLM was compared with other methods (LTMG, CSN, Seurat-wgcna, and MNN-wgcna) on four synthetic cohorts. Each synthetic cohort contains 9 sets of simulated data with an increasing number of samples. The bar plot represents the ARI of identified gene clusters compared to the ground truth. **A.** ARI of synthetic cohort 1. **B.** ARI of synthetic cohort 2. **C.** ARI of synthetic cohort 3. **D.** ARI of synthetic cohort 4. ARI, Adjusted Rand Index; LTMG, Left Truncated Mixture Gaussian; CSN, Cell Specific Network.

s41467-018–06052-0 [23]. For our analysis, we used malignant cells and genes expressed in over 300 cells as input.

Clustering evaluation index

Each clustering result produced by different methods was assessed using clustering evaluation indices, including the Adjusted Rand Index (ARI) [43], the Calinski-Harabasz (CH) index [44], the Davies-Bouldin (DB) index [45], and the Dunn index [46]. CH index evaluated the cluster validity based on the average between- and within-cluster sum of squares. DB index was obtained by averaging all the cluster similarities. Dunn index used the minimum pairwise distance between objects in different clusters as the inter-cluster separation and the maximum diameter among all clusters as the intra-cluster compactness. Larger CH index, smaller DB index, and larger Dunn index represented better clustering results.

Cell clustering based on co-expressed gene modules

With the co-expressed gene modules, we utilized mean value of the modules in each single cell as input for graph-based clustering. Uniform manifold approximation and projection (UMAP) was used to visualize cell clusters. Graph-based clustering was performed using the Seurat package (v3.1), and UMAP analysis was performed using the "umap" package (v.0.2.3.1) [47] in R (v.3.4.3). The number of epochs (n_epochs) was set at 20. The n_neighbors value was set at 15, and min_dist was set as 0.1.

Statistical analysis

Kaplan-Meier (KM) analysis was performed using the "survival" R package (http://cran.r-project.org/web/packages/survival/index.html). Log-rank test was used to test the

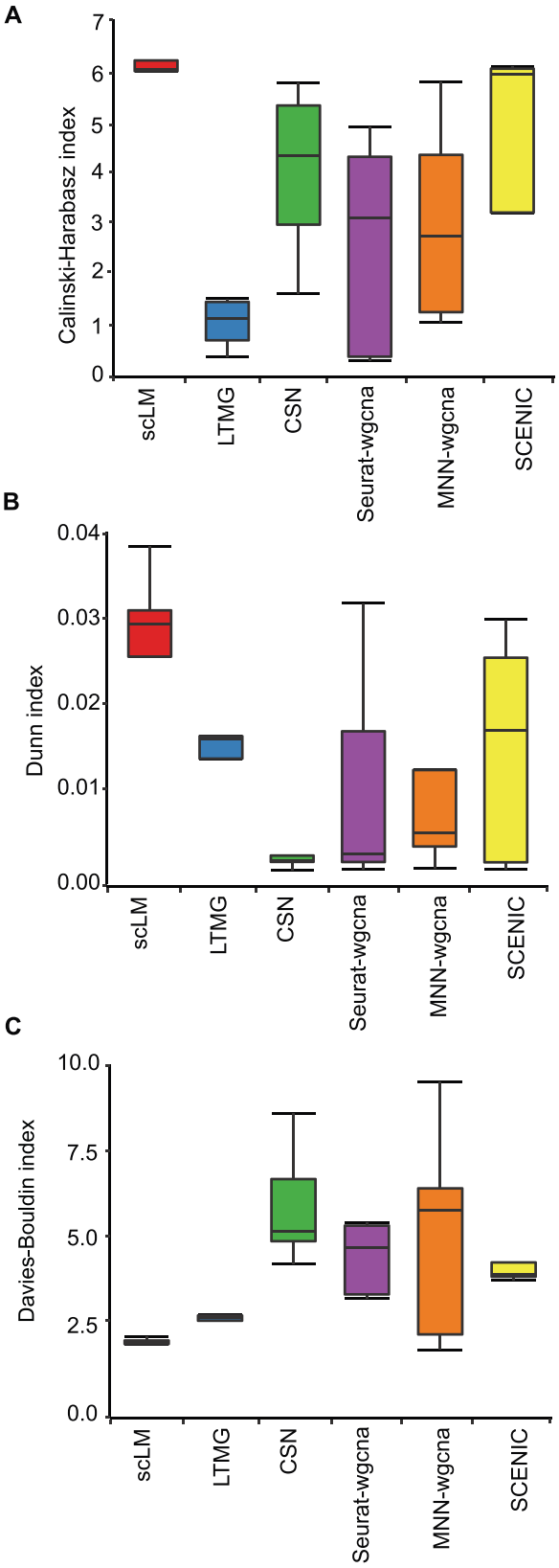

Figure 3 Evaluation of scLM using experimental data
scLM was compared with other methods (LTMG, CSN, Seurat-wgcna, MNN-wgcna, and SCENIC) on five experimental datasets. Multiple evaluation indices were used, including: the Calinski-Harabasz index (**A**), the Dunn index (**B**), and the Davies-Bouldin index (**C**).

differences of survival curves. When evaluating the performance of scLM, P value was calculated by t-test.

Functional analysis

Hallmark collection

We downloaded the Hallmark gene set collection for functional analyses from Molecular Signatures Database (MSigDB) [48], which was a widely used and comprehensive database. Each hallmark in this collection consisted of a "refined" gene set that conveyed a specific biological state or process and displayed coherent expression. The hallmarks effectively summarized most of the relevant information of the original founder sets and, by reducing both variation and redundancy, provided more refined and concise inputs for gene set enrichment analysis.

Pathway database

Reactome (http://www.reactome.org) was a manually curated open-data resource of human pathways and reactions, which was an archive of biological processes and a tool for discovering potential functions. Gene sets derived from the Reactome [49] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [50] pathway database were downloaded from the MSigDB Collections.

Enrichment test

Functional enrichment based on the Reactome and Gene Ontology (GO) databases was assessed by hypergeometric test, which was used to identify a priori-defined gene set that showed statistically significant differences between two given clusters. Enrichment test was performed by the clusterProfiler package [51]. Test P values were further adjusted by Benjamini-Hochberg correction, and adjusted P values less than 0.05 were considered statistically significant.

Results and discussion

Overview of scLM

We developed a new method, scLM, for simultaneously identifying consensus co-expressed genes from multiple scRNA-seq datasets. Our hypothesis was that co-expressed genes coordinating biological processes could be captured across multiple different datasets. In our model, we assumed that latent variables captured the intrinsic signals of the co-expressed genes regardless of technical variances and batch effects among different datasets. Figure 1 provided an illustrative overview of the scLM method. Briefly, the input contained a collection of multiple datasets (k) representing the single-cell sequencing data generated under different clinical or experimental settings. In the k-th dataset, we assumed that the observed expression levels, x_{ijk}, of the i-th gene across cells $j \in \{1, \cdots, n_k\}$ followed the NB distribution $NB(u_{ik}, \theta_{ik})$. The intrinsic biological variability of gene i across all cells and all datasets was captured by the latent variables z_i in a λ-dimension latent space. This was achieved through a conditional GLM log $u(x_{ijk}|z_i) = \alpha_{jk} + \beta_{jk} z_i$ that distinguished the intrinsic biological variability z_i from the extrinsic signals (α_{jk} and β_{jk}) including the technical variances at the

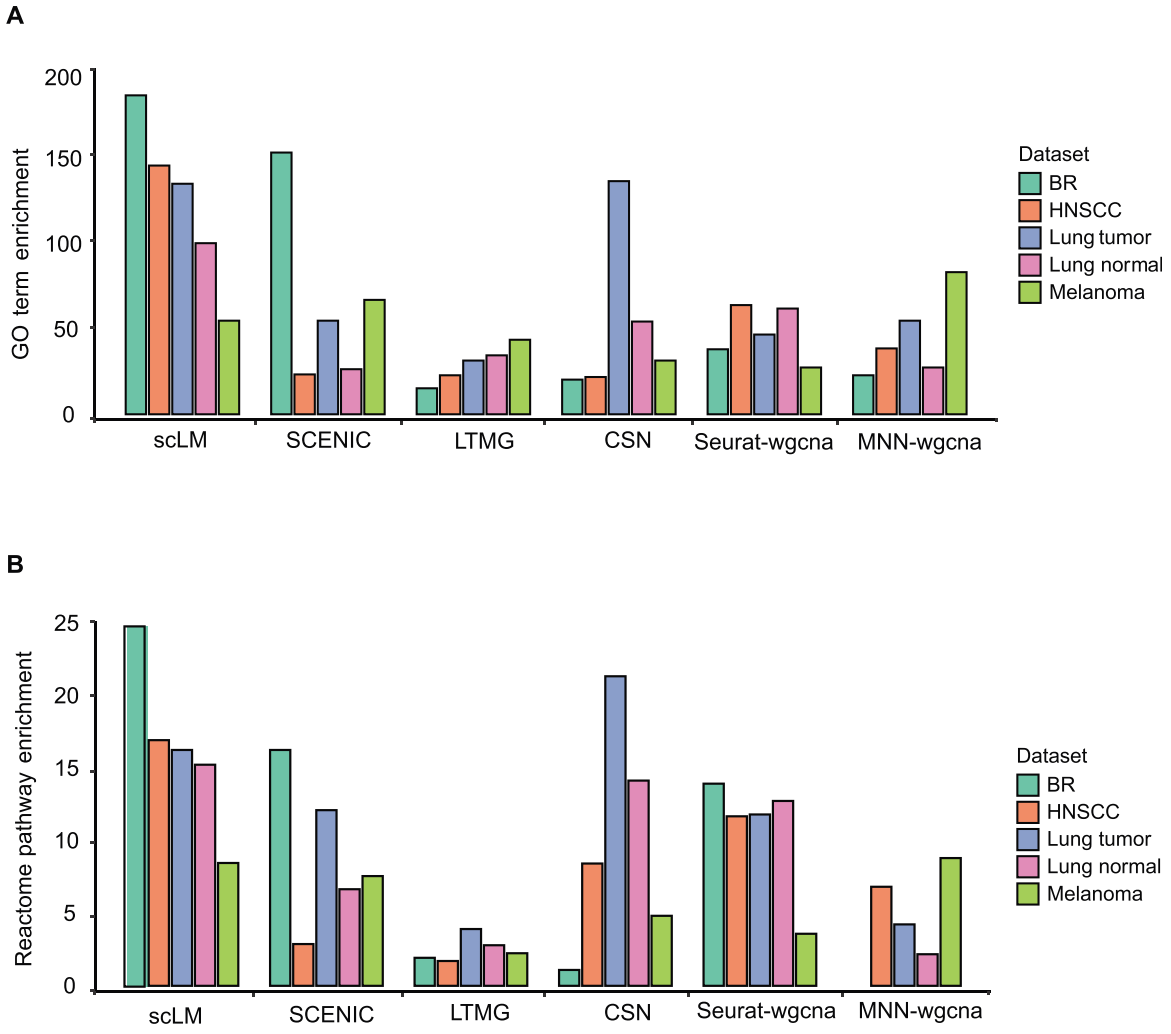

Figure 4 scLM identifies co-expressed genes with significantly enriched biological functions

A. The average number of significantly enriched GO terms (adjusted P value < 0.05) based on the co-expressed genes identified by different methods. **B.** The average number of significantly enriched pathways in Reactome database (adjusted P value < 0.05) based on the co-expressed genes identified by different methods. GO, Gene Ontology.

cell level (j) and batch effects at the sample level (k). The latent variables and other parameters were estimated and obtained using MCMC and maximum likelihood approaches. Therefore, different groups of co-expressed genes (m1–m4) across multiple datasets were identified through clustering genes in the latent space. Further explanations of the mathematical model were included in the Method.

Performance evaluation on simulation data

To evaluate the performance of scLM, we benchmarked it against other methods, including LTMG [32], CSN [31], Seurat-wgcna, MNN-wgcna, and SCENIC [30]. Seurat-wgcna and MNN-wgcna referred to the co-expression analysis using WGCNA [28], following the batch correction by Seurat [52] or MNN [53]. As SCENIC relied on the RcisTarget database that required real gene input, we omitted comparing with SCENIC on simulation data but still included it in the comparison on real single-cell data.

We first generated two synthetic data cohorts (synthetic cohorts 1 and 2) from NB distribution. Each cohort contained 9 sets (D1–D9) of simulated data with an increasing number of samples. That is, D1 contained one individual dataset ($n = 1$), D2 contained two individuals of datasets ($n = 2$), and so on. Each set contained three co-expressed gene clusters as ground truth. Additionally, we utilized the Splatter package [40] to generate another two batches of simulated data (synthetic cohorts 3 and 4) with dropout effects, which could more accurately recapitulate actual scRNA-seq data distributions. Details of the simulation datasets were provided in the Method.

With the simulated data cohorts, we applied scLM and other methods (LTMG, CSN, Seurat-wgcna, and MNN-wgcna) to identify the co-expression clusters. To assess and quantify clustering accuracy, we used the ARI [43] as the performance metric to rank these methods (**Figure 2**). The corresponding bar plots represented the ARI of the identified clusters by each method compared to the ground truth. Notably, scLM accurately identified each gene cluster in four

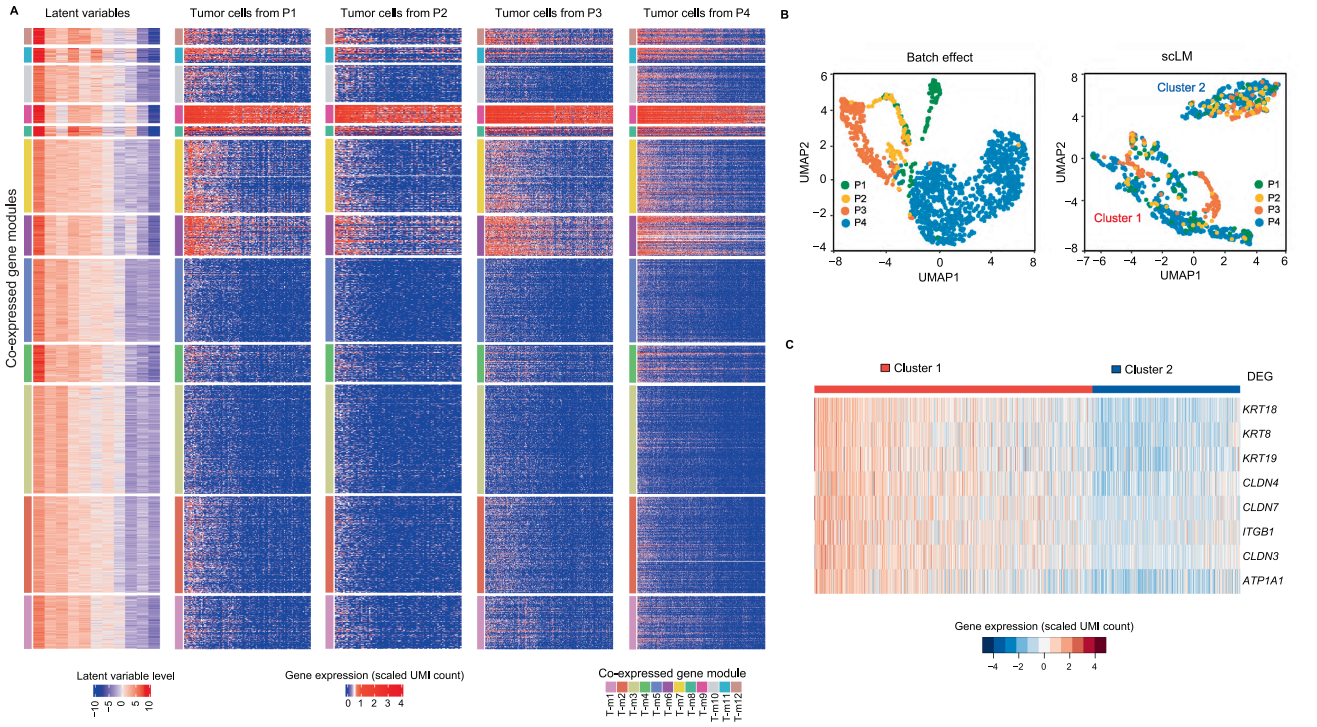

Figure 5 scLM identifies co-expressed gene modules that characterize subtle cell subpopulations

A. Simultaneous and consensus clustering of genes across lung tumor cells from four patients (P1–P4). scLM reveals 12 co-expressed gene modules characterized by the latent variables as well as shown in gene expression data of four patients. In each heatmap, rows are genes assigned to 12 modules. In each co-expressed gene module, genes are consistently over-expressed (red) or under-expressed (blue). **B.** The left panel shows strong batch effects of different patients. The right panel depicts the UMAP visualization of single cells characterized by the co-expressed gene modules. Two evenly distributed clusters (cluster 1 and cluster 2) are identified. Different patients are distinguished by colors. **C.** Heatmap shows the differential expression pattern of EMT-related genes between two clusters. UMI, unique molecular identifier; UMAP, uniform manifold approximation and projection; DEG, differentially expressed gene.

cohorts, and demonstrated much higher ARIs (mean ± SE: 0.979 ± 0.063 for synthetic cohort 1; 0.971 ± 0.031 for synthetic cohort 2; 0.899 ± 0.043 for synthetic cohort 3; 0.886 ± 0.025 for synthetic cohort 4). The other methods showed relatively lower ARIs. For example, CSN showed lower ARIs in synthetic cohort 1 (mean ± SE: 0.627 ± 0.028) and synthetic cohort 3 (mean ± SE: 0.520 ± 0.070). LTMG presented little higher ARIs and lower variances in four synthetic cohorts. These results demonstrated the outperformance of scLM in identifying accurate co-expressed genes from multiple datasets.

Evaluation of scLM using experimental data

To further demonstrate the performance of scLM, we compared scLM with other methods (LTMG, CSN, Seurat-wgnca, MNN-wgcna, and SCENIC) on experimental scRNA-seq datasets. For comparisons, we used two in-house datasets from lung tumor and adjacent normal tissues as well as three public datasets from BR, HNSCC, and melanoma. The data pre-processing procedures were described in the Method.

To assess and quantify clustering accuracy on real datasets, we used performance metrics including the CH index [44], Dunn index [46], and DB index [45], to rank these methods. Importantly, scLM produced sets of clusters that showed significantly higher CH values than other methods (**Figure 3**A), especially higher than LTMG (P = 1.75E−07) and MNN-wgcna (P = 0.02), demonstrating that scLM achieved better cluster validity than other methods based on average between- and within-cluster sum of squares. In addition, compared to other methods, scLM also achieved significantly higher Dunn index scores representing better inter-cluster separation and intra-cluster compactness (Figure 3B), and lower DB index scores reflecting higher cluster quality (Figure 3C). Though SCENIC and Seurat-wgcna showed higher Dunn index score in one dataset (HNSCC), they failed to show superior performance on other datasets. Thus scLM proved to achieve the best partitioning of co-expressed gene clusters that are most distinct from each other.

scLM identified co-expressed genes with significantly enriched biological functions

As co-expressed genes were likely to be enriched with biological functions, we compared the extent to which different methods affected the functional discovery, based on their identified co-expressed genes. First, the aforementioned methods were evaluated for their capability to detect enriched GO terms in the five experimental datasets. Different methods identified gene clusters enriched with different GO enrichment results. The average number of significantly enriched GO terms (adjusted P value < 0.05) ranged from 15 to 184 (**Figure 4**A).

scLM extracted co-expressed genes with more enriched functional terms than other methods in three of the five datasets (*i.e.*, the BR, HNSCC, and lung Normal datasets). SCENIC identified relatively high number of enriched GO terms in the BR dataset, whereas low number of enriched GO terms in other datasets; CSN identified relatively high number of enriched GO terms in the lung Tumor dataset but low number of enriched GO terms in other datasets; MNN-wgcna identified relatively high number of enriched GO terms in the melanoma dataset but low number of enriched GO terms in other datasets. LTMG and Seurat-wgcna showed lower number of enriched GO terms in five datasets. Similar results were observed when we strengthened the enriched significance by the adjusted P value < 0.01 (Figure S1). The number of significant terms became fewer for all the methods, yet scLM identified the most on all datasets except for the melanoma dataset. Some methods, like LTMG, failed to identify gene clusters with enriched terms at the threshold of adjusted P value < 0.01.

In addition to GO terms, we also examined the enriched pathways in the Reactome database, based on the co-expressed genes identified by different methods (Figure 4B). Different methods showed different pathway enrichment results. Importantly, scLM identified co-expressed genes with more enriched pathways than other methods in three of the five datasets (*i.e.*, the BR, HNSCC, and lung Normal datasets). Taken together, these results demonstrated that scLM outperforms other methods in functional discovery of co-expressed genes.

scLM identified the tumor-specific modules enriched in specific cell state

In real-world scenarios, samples from different patients or different data sources often demonstrated highly different cell numbers, largely due to strong batch effects and technical issues. The scLM method was designed to address such highly unbalanced data that outperformed other competitors on such datasets. To validate the effectiveness of scLM, we intentionally selected patient samples that varied with respect to cell number, which could create challenges for this method. As a case study, we used scLM to analyze our in-house scRNA-seq profiling from 4 NSCLC patients (P1–P4) [41] to identify the co-expressed genes in tumor and normal epithelial cells, respectively. In tumor cells (**Figure 5**A, heatmap of latent variables), we discovered 12 co-expressed gene modules in the latent space (T-m1–T-m12). These modules showed clear differences but were consistently concordant across patients (Figure 5A, heatmaps of P1–P4), even though the single cells from different patients presented strong heterogeneity and batch effects (Figure 5B, left panel).

Using the 12 co-expression modules, the single cells were separated into two major clusters. In each cluster, cells from different patients mixed well without interference from batch effects (Figure 5B, right panel), which further supported that the co-expression modules were consistent across patients and not affected by batch effects. Interestingly, we found that cluster 1 had higher expression of epithelial functional markers (EMT-related genes) than cluster 2 (Figure 5C). These results indicated that co-expression modules were capable of characterizing specific cell phenotypes.

Similarly, in normal single cells, we observed 13 co-expressed gene modules (N-m1–N-m13) that showed concordant expression across individual patients (Figure S2). Then we compared the co-expressed gene modules identified from tumor and normal cells. Four modules (T-m1, T-m3, T-m4, and T-m10) were not correlated with any normal modules, suggesting that they were tumor-specific (File S1; Figure S3).

scLM identified a common program across three types of cancer

To explore the underlying mechanism of carcinogenesis, we next extended the application of scLM to HNSCC and melanoma. In addition to the 12 co-expressed gene modules identified in NSCLC, we identified 11 modules in HNSCC and 14 modules in melanoma. To determine the similarities of these co-expression modules, we performed a pair-wise comparison using weighted Jaccard similarity, followed by hierarchical clustering. As shown in the diagram (Figure S4A), we found that most branches were dominated by a mixture of cancer types. Importantly, we identified a branch with high similarity among T-m9, HNSCC-m7, and Melanoma-m12 modules.

These three similar modules substantially overlapped with 91 genes, which were defined as a common program across three cancer types. To gain insights into the biological functions of the common program, we performed enrichment analysis in the Hallmark database (Figure S4B). The *MYC* targets v1 and hypoxia were the top enriched terms, involving the genes *FOS*, *GAPDH*, *HLA-A*, and *NFKB1A*, which suggested the common oncogenesis mechanism regardless of cancer types (File S1).

From the applications of scLM, we see three meaningful use of scLM to scientific research. 1) scLM identifies co-expressed genes that reveal novel biological processes. An example is the lung tumor-specific module that highlights cell–cell communication in tumor microenvironment (File S1; Figure S3). 2) scLM contributes to the subtle characterization of cell states. In lung cancer, scLM identifies 12 co-expression modules that are consistent across patients. These co-expression modules separate cells into two major clusters, of which one cluster presents different EMT activity suggesting more precise characterization of cell states (Figure 5). 3) With the co-expressed genes identified by scLM, both specific and common gene modules can further be explored for their translational and biological relevance. For example, in melanoma, scLM identifies two co-expressed gene modules that are associated with immune checkpoint inhibitor (ICI) resistance, which provides potential value for predicting ICI therapy response. We also find a common co-expression module from three different cancer types, and reveal the *MYC* targets and hypoxia as the common intrinsic mechanisms of tumor malignancy (File S1; Figure S4).

Given the merits of scLM, several potential limitations warrant further study. First, zero-inflated genes are excluded during pre-processing. The main reason is that, genes with inflated zeros are not informative and have negligible meaningful contribution to co-expression. The other reason is, with the fast advance of scRNA-seq technology, zero-inflation issue will be very minimal in near future. Second, in future work, we will examine the necessity of providing zero-inflation models, which specifically deal with data of poor sequencing depth and strong dropout effects. Third, the computational cost of

scLM can be further reduced. We have already utilized C programming and parallel computing to dramatically boost the efficiency of scLM. However, considering that scRNA-seq data are growing into million-cell level, we will explore the use of GPU computing and cloud-based approaches to catch up with the scale of future scRNA-seq data.

Conclusion

Co-expressed genes with coordinate expression indicate functional linkages between genes. Genes with coordinate biological functions are frequently co-transcribed, resulting in co-expression profiles. Thus, co-expressed genes can be used to intuitively associate genes with biological processes, to reveal disease-related genes, and to discern transcriptional regulatory mechanisms. Accumulative evidence supports the reliability of co-expression analysis for annotating and inferring gene functions [1–3,9–12]. Recent advances in scRNA-seq technologies enable the systematic interrogation of gene co-expression modules in specific cell types, and the elucidation of the underlying biological mechanisms [13,54,55]. The improved data resolution and quality allow accurate identification of disease-related modules and regulatory genes for specific cell types and specific tissues. Thus, we expect co-expression analysis to be more widely applied due to the technology advances.

In this study, we introduce a novel method, scLM, to simultaneously identify co-expressed genes across multiple single-cell datasets. The scLM algorithm uses the conditional NB distribution with latent variables to disentangle co-expression patterns across multiple datasets. To our knowledge, scLM is the first available tool that is capable of leveraging multiple scRNA-seq datasets to accurately detect co-expressed genes. We provide an overview of scLM and illustrate how scLM can be used to further characterize cell states and identify tumor-specific modules in lung cancer. We demonstrate that the tumor-specific modules are enriched in pathways, including cell–cell communication and *SMAD2/3/4* transcriptional activity, with identified upstream transcriptional factors including *TEAD1* and *FOXA1*. We further show the clinical prognostic significance of these discoveries in clinical samples. Moreover, we explore the common co-expressed genes, *i.e.*, the common module, across three cancer types and offer intrinsic mechanisms of tumor malignancy. The common module is highly enriched in the *MYC* targets v1 and hypoxia, suggesting the presence of common intrinsic oncogenesis mechanisms. Additionally, the common module is shown to be related with clinical response to ICI in melanoma patients, suggesting that the common module provides predictive value of ICI therapy response.

Compared with other methods, scLM has several key advantages: 1) scLM accounts for data heterogeneity and variances among multiple datasets, such as unbalanced sequencing depths and technical biases in library preparation. 2) scLM leverages information across datasets for detecting stable and conserved co-expression modules with high accuracy and reproducibility. 3) scLM is an integrated pipeline that uses raw count matrix without prior batch-correction as input, thus can be easily applied to scRNA-seq data. Overall, scLM opens possibilities for further investigation and mechanistic interpretation of co-expressed genes. With the growing scRNA-seq data, scLM is poised to become a valuable tool for elucidating co-expression studies in single-cell transcriptomics.

Ethical statement

For our in-house data, fresh remnant tumor and adjacent normal tissues were collected at the time of elective curative resection by the Tumor Tissue and Pathology Shared Resource (TTPSR) of the Wake Forest Baptist Medical Center Comprehensive Cancer Center (WFBMC-CCC), USA. Collections by the TTPSR adhere to Institutional Review Board approved procedures (Advanced Tumor Bank protocol CCCWFU 01403, TTPSR collections IRB BG04-104 which also allows for the use of de-identified protected health information along with the tissue samples). Acquisition of de-identified samples from the TTPSR for single-cell isolation and research use was in accordance with approved IRB protocol 00048977. All patients provided written consent to participate in the study, which was approved by the institutional reviewing board.

Code availability

A user-friendly R package with all the key features of the scLM method is available at https://github.com/QSong-github/scLM.

CRediT author statement

Qianqian Song: Data curation, Methodology, Software, Writing - original draft. **Jing Su:** Methodology, Software, Validation, Writing - review & editing. **Lance D. Miller:** Investigation, Writing - review & editing. **Wei Zhang:** Conceptualization, Supervision. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

The authors thank Elizabeth Forbes and Karen Klein for professional editing service. This work was supported in part by the Cancer Genomics, Tumor Tissue Repository, and Bioinformatics Shared Resources under the NCI Cancer Center Support Grant to the Comprehensive Cancer Center of Wake Forest University Health Sciences, USA (Grant No. P30CA012197). WZ is supported by the Hanes and Willis Professorship in Cancer, USA. Additional support for QS and WZ are provided by a Fellowship to WZ from the National Foundation for Cancer Research, USA. JS is supported by Indiana University Precision Health Initiative, USA.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.09.002.

ORCID

0000-0002-4455-5302 (Qianqian Song)
0000-0003-4917-6173 (Jing Su)
0000-0003-3799-2528 (Lance D. Miller)
0000-0002-2235-1851 (Wei Zhang)

References

[1] Ferrari R, Forabosco P, Vandrovcova J, Warren JD, Momeni P, Weale ME, et al. Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. Mol Neurodegener 2016;11:21.
[2] Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun 2014;5:3231.
[3] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.
[4] Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell 2018;175:984–97.e24.
[5] Singer M, Wang C, Cong L, Marjanovic ND, Kowalczyk MS, Zhang H, et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. Cell 2017;171:1221–3.
[6] Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 2017;171:1611–24.e24.
[7] Chihara N, Madi A, Kondo T, Zhang H, Acharya N, Singer M, et al. Induction and transcriptional regulation of the co-inhibitory gene module in T cells. Nature 2018;558:454–9.
[8] Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature 2015;526:131–5.
[9] Stäehler CF, Keller A, Leidinger P, Backes C, Chandran A, Wischhusen J, et al. Whole miRNome-wide differential co-expression of microRNAs. Genomics Proteomics Bioinformatics 2012;10:285–94.
[10] Clements M, van Someren EP, Knijnenburg TA, Reinders MJT. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. Genomics Proteomics Bioinformatics 2007;5:86–101.
[11] Zheng CH, Huang DS, Kong XZ, Zhao XM. Gene expression data classification using consensus independent component analysis. Genomics Proteomics Bioinformatics 2008;6:74–82.
[12] Wan P, Wu J, Zhou Y, Xiao J, Feng J, Zhao W, et al. Computational analysis of drought stress-associated miRNAs and miRNA co-regulation network in *physcomitrella patens*. Genomics Proteomics Bioinformatics 2011;9:37–44.
[13] Xhangolli I, Dura B, Lee G, Kim D, Xiao Y, Fan R. Single-cell analysis of CAR-T cell activation reveals a mixed TH1/TH2 response independent of differentiation. Genomics Proteomics Bioinformatics 2019;17:129–39.
[14] Yu P, Lin W. Single-cell transcriptome study as big data. Genomics Proteomics Bioinformatics 2016;14:21–30.
[15] Svensson V. Droplet scRNA-seq is not zero-inflated. Nat Biotechnol 2020;38:147–50.
[16] Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016;17:75.
[17] Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics 2017;33:3486–8.
[18] Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods 2014;11:637–40.
[19] Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol 2016;17:63.
[20] Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res 2014;24:496–510.
[21] Kolodziejczyk A, Kim JK, Svensson V, Marioni J, Teichmann S. The technology and biology of single-cell RNA sequencing. Mol Cell 2015;58:610–20.
[22] Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. PLoS Genet 2014;10:e1004126.
[23] Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 2018;174:1293–308.e36.
[24] Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of *in vivo* mammalian chromatin accessibility. Cell 2018;174:1309–24.e18.
[25] Muraro MJ, Dharmadhikari G, Grun D, Groen N, Dielen T, Jansen E, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst 2016;3:385–94.e3.
[26] Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. Nature 2018;562:367–72..
[27] Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell 2018;173:1535–48.e16.
[28] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
[29] Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biol 2018;19:172.
[30] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14:1083–6.
[31] Dai H, Li L, Zeng T, Chen L. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res 2019;47:e62.
[32] Wan C, Chang W, Zhang Y, Shah F, Lu X, Zang Y, et al. LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Res 2019;47:e111.
[33] Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol 2006;4:e309.
[34] Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 2013;10:1093–5.
[35] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
[36] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res 2012;40:4288–97.
[37] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

[38] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 2019;20:1–15.
[39] Wang Z, Ma S, Zappitelli M, Parikh C, Wang CY, Devarajan P. Penalized count data regression with application to hospital stay after pediatric cardiac surgery. Stat Methods Med Res 2016;25:2685–703.
[40] Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 2017;18:174.
[41] Song Q, Hawkins GA, Wudel L, Chou PC, Forbes E, Pullikuth AK, et al. Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. Cancer Med 2019;8:3072–85.
[42] Tirosh I, Izar B, Prakadan SM, Wadsworth 2nd MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352:189–96.
[43] Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.
[44] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat 1974;3:1–27.
[45] Davies DL, Bouldin DW. A cluster separation measure. IEEE PAMI 1979;2:224–7.
[46] Dunn† JC. Well-separated clusters and optimal fuzzy partitions. J Cybernetics 1974;4:95–104.
[47] Krijthe JJCS. Rtsne: T-distributed stochastic neighbor embedding using a Barnes-Hut implementation. R package version 0.13, 2015. https://github. com/jkrijthe/Rtsne.
[48] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011;27:1739–40.
[49] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome pathway knowledgebase. Nucleic Acids Res 2018;46:649–55.
[50] Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol Biosyst 2014;10:2441–7.
[51] Yu GC, Wang LG, Han YY, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[52] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.e21.
[53] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421–7.
[54] Wang D, Gu J. VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Genomics Proteomics Bioinformatics 2018;16:320–31.
[55] Ren X, Zheng L, Zhang Z. SSCC: a novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. Genomics Proteomics Bioinformatics 2019;17:201–10.