FEM: mining biological meaning from cell level in single-cell RNA sequencing data (#60379)

First submission

Guidance from your Editor

Please submit by **1 Jun 2021** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

6 Figure file(s)
7 Table file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Speculation is welcome, but should be identified as such.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

FEM: mining biological meaning from cell level in single-cell RNA sequencing data

Yunqing Liu [1], **Na Lu** [1], **Changwei Bi** [1], **Tingyu Han** [1], **Zhuojun Guo** [1], **Yunchi Zhu** [1], **Yixin Li** [1], **Chunpeng He** [Corresp., 1], **Zuhong Lu** [Corresp. 1]

[1] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering,, Southeast University, Nanjin, Jiangsu, China

Corresponding Authors: Chunpeng He, Zuhong Lu
Email address: cphe@seu.edu.cn, zhlu@seu.edu.cn

Background. One goal of expression data analysis is to discover the biological significance (or function) of genes that are differentially expressed. As one of the main tools for function mining, Gene Set Enrichment (GSE) analysis has been widely used. However, for single-cell RNA sequencing (scRNA-SEQ) data, every gene expressed in a cell is valuable information for GSE and not should be discarded. **Methods.** To utilize the information of all expressed genes, we developed the FEM algorithm, which converts the gene expression matrix (GEM) into a functional expression matrix (FEM). The FEM algorithm can not only explain the biological significance of a single cell but also can be used to replace or integrate GEM for downstream analysis. **Results.** Applying FEM to the three datasets (PBMC, human liver, and human pancreas), we found that FEM showed good performance in cell clustering, and cell type specified function annotation.

FEM: mining biological meaning from cell level in Single-cell RNA sequencing data

Yunqing Liu[1], Na Lu[1], Changwei Bi[1], Tingyu Han[1], Zhuojun Guo[1], Yunchi Zhu[1], Yixin Li[1], Chunpeng He[1], Zuhong Lu[1]

[1] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096
Corresponding Author:
Chunpeng He[1], Zuhong Lu[1]
Sipailou 2, Nanjing, Jiangsu Province, 210096, China
Email address: zhlu@seu.edu.cn

Abstract

Background. One goal of expression data analysis is to discover the biological significance (or function) of genes that are differentially expressed. As one of the main tools for function mining, Gene Set Enrichment (GSE) analysis has been widely used. However, for single-cell RNA sequencing (scRNA-SEQ) data, every gene expressed in a cell is valuable information for GSE and not should be discarded.

Methods. To utilize the information of all expressed genes, we developed the FEM algorithm, which converts the gene expression matrix (GEM) into a functional expression matrix (FEM). The FEM algorithm can not only explain the biological significance of a single cell but also can be used to replace or integrate GEM for downstream analysis.

Results. Applying FEM to the three datasets (PBMC, human liver, and human pancreas), we found that FEM showed good performance in cell clustering, and cell type specified function annotation.

Introduction

As an alternative to the microarray platform, RNA sequencing (RNA-SEQ) has been widely used in the past 10 years. It has provided many valuable insights into the complex biological mechanism, ranging from cancer genomics to diverse microbial communities. Compared with traditional bulk methods that profile batches of cells in a pooled way, single-cell RNA-SEQ (scRNA-SEQ) can facilitate new and potentially unexpected biological discoveries. For example, it can provide information on complex and rare cell populations, regulatory relationships between genes, and can track the trajectories of distinct cell lineages during development (Gawad, Koh & Quake, 2016) (Hwang, Lee & Bang, 2018) (Chen, Ning & Shi, 2019).

The cell is considered to be the fundamental unit in biology. For centuries, biologists have known that multicellular organisms are characterized by a plethora of distinct cell types. Because of homogeneity in the genome, the difference between cells in organisms can be characterized based on transcriptome similarity, which can be defined through unsupervised clustering (Kiselev, Andrews & Hemberg, 2019). Transcriptomic studies using bulk tissue assume that all cells from a homogeneous material, thereby ignoring the cellular heterogeneity of the sample. Single-cell transcriptomics, however, can address the above question (Kulkarni et al., 2019). High-throughput single-cell transcriptomics has provided unprecedented insights into cellular diversity in tissues across diverse organisms. scRNA-SEQ is a promising approach for the study of the transcriptomes of individual cells in organisms.

However, the pipeline for scRNA-SEQ is still based on bulk RNA-SEQ. The basic workflow includes data pre-processing (quality control (Luecken & Theis, 2019), (Ilicic et al., 2016), (Griffiths, Scialdone & Marioni, 2018), normalization, data correction, feature selection, and dimensionality reduction), followed by cell-level and gene-level downstream analysis (Luecken & Theis, 2019). Because RNA-SEQ data comprise pooled samples of multiple cells, many genes are expressed. But it is unclear whether these genes are co-expressed or if this is just the result of

mixing. Therefore, a portion of the particularly well-characterized genes needs to be screened for downstream analysis.

For scRNA-SEQ data, this analysis process has the following shortcomings. First, the characteristics of single-cell data are not fully utilized: scRNA-SEQ data are a direct reflection of the physiological state of a cell but the above method is an analytical process based on multiple cells or cell clusters. Second, it cannot effectively capture all meaningful functional groups because it filters many genes that may play important roles in transient cell states. Since functional enrichment analysis is based on all the expressed genes in a cell, including genes that were filtered in a previous process, it is possible to miss meaningful biological functions. Third, the results of downstream functional analysis and upstream clustering are not been integrated and visualized. To overcome these limitations, this study proposes a method based on transforming the gene expression matrix into a functional expression matrix.

Materials & Methods

Motivation for this approach

Functional Gene Set Enrichment Analysis (GSEA) is usually the last step of expression data analysis. A gene set usually represents a biological function, and the function will be used to refer to a gene set in the following. There are many R packages and some online sites such as DAVID that provide enrichment analysis tools. The software takes a set of genes given by the user as input and returns the functions (such as pathways) that are significantly enriched. Enrichment analysis methods for individual samples have also been proposed(Foroutan et al., 2018)(Hänzelmann, Castelo & Guinney, 2013).

Because of the low amount of initial material, scRNA-SEQ has limitations in low capture efficiency and high dropouts. scRNA-SEQ expression data tend to be sparse(Hicks et al., 2018b) (Eraslan et al., 2019). Due to the non-trivial distinction between true- and false-zero counts, the true zero represents the lack of expression of a gene in a specific cell, while a false zero is a technical deviation. Because the RNA-SEQ technique only detects mRNA molecules that are present, a gene in the scRNA-SEQ dataset with a non-zero expression value means that at least one mRNA molecule is present. These sparse non-zero expression values provided inspiration for the possibility of performing functional GSEA analysis at the single-cell level.

Given the characteristics of scRNA-SEQ data, functional enrichment analysis can be performed at the single-cell transcriptome expression level first, replacing the initial expression matrix with a functional expression matrix (FEM) can allow the direct exploration of the differences between cells from different biological perspectives. Second, combined with the gene expression matrix (GEM), the top variable function of a cell cluster obtained by GEM can be represented as a cell scatter plot. Third, for GSEA, the coverage of a functional gene set is more important than the expression of individual genes, because genes often overlap different functional groups, which makes single high-expression genes hard to explain. For example, the activation of a signalling pathway is the result of interactions between all genes; high expression of one gene does not mean that the pathway is activated.

At present, the main factors affecting the discovery of cell groups are as follows. First, technical covariates must be regressed out before downstream analysis as these factors will introduce systematic error and confound the technical and biological variability, leading to systematic differences in gene expression profiles between batches(Leek et al., 2010) (Hicks et al., 2018a) (Chen, Ning & Shi, 2019). The most prominent technical covariates in single-cell data are *count depth* and *batch*. The FEM method is based on a gene set, which makes it more robust than the GEM. Second, some biological effects can affect the results of the cluster algorithms. For example, the cell cycle can alter the clustering result in non-proliferating cell populations. However, correcting for biological covariates is not always helpful in interpreting scRNA-seq data (Kolodziejczyk et al., 2015). These influencing factors often do not have uniform filtering criteria. In some cases, the cell cycle may be part of the study, or there may be a relationship between the cell cycle and other functions (Haghverdi, Buettner & Theis, 2015) (Vento-Tormo et al., 2018) (McDavid, Finak & Gottardo, 2016) (Blasi et al., 2017). The FEM method can be used to systematically survey the biological aspect of each cell before downstream analysis. To this end, we developed a scRNA-SEQ functional expression matrix algorithm (FEM).

Workflow of the proposed methods

The FEM was divided into four steps (***Fig. 1***). **1.** The standardized scRNA-SEQ GEM was transformed into a FEM by multi-module gene enrichment analysis (gene ontology, pathway). **2.** Because the *p*-value represents the significance of enrichment, the *p*-value obtained by enrichment analysis was converted into information content (FEM). **3.** FEM was used instead of GEM for data standardization, dimensionality reduction clustering, and UMAP(McInnes et al., 2018) visualization. **4.** The FEM and GEM were integrated to find differentially expressed genes (DEG) and differentially expressed functions (DEF).

Dataset

Peripheral blood mononuclear cells (PBMCs) are populations of immune cells that remain at the less dense upper interface of the Ficoll layer. PBMCs include lymphocytes (T cells, B cells, and NK cells), monocytes, and dendritic cells (DCs). In humans, the frequencies of these populations vary across individuals. Lymphocytes are typically in the range of 70–90%, monocytes range from 10–30%, while DCs are only present at 1–2% (NORMAN, 1995). The PBMC dataset (Butler et al., 2018) used in this paper (downloaded from the official Seurat site) mainly included B cells, NK cells, CD8 T cells, memory CD4 T cells, Naïve CD4 T cells, DC, CD14+ monocytes, FCGR3A+ monocytes, and a small number of platelets. This dataset contains 2,700 cells.

The human pancreas dataset contains 2,126 cells and 10 cell types. It mainly includes alpha cells, ductal cells, endothelial cells, delta cells, acinar cells, beta cells, gamma cells, mesenchymal cells, epsilon cells, and a small number of unknown types of cells(Muraro et al., 2016).

The human liver dataset consists of 777 cells, mainly including seven types of cells: definitive endoderm cells, immediate hepatoblast cells, induced pluripotent stem cells (IPSCs), material

hepatocytic cells, hepatic endoderm cells, endothelial cells, and mesenchymal stem cells(Camp et al., 2017).

FEM algorithm

Selected functional groups and their profiles

Three functional gene sets were selected from the Msigdb database for enrichment analysis, the Reactome pathways, gene ontology (GO), and immunologic signature gene set (Liberzon et al., 2011) (Table 1).

Gene-functional group conversion

The non-zero–expressed genes of the cells in the GEM were extracted first. In the second step an enrichment analysis score for each cell was calculated. The enrichment analysis method was based on Fisher's exact test using the Python SciPy package. The Fisher exact test is a statistical test based on a hypergeometric distribution that is used to determine if there are non-random associations between two categorical variables, or to test whether the theoretical value is consistent with the actual value.

$$P = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (1)$$

Here, N represents the total number of background genes, K represents the number of genes in a particular gene set, n represents the number of non-zero genes in a single cell, and k represents the number of genes present in both K and n. The Bonferroni correction was used to counteract the problem of multiple comparisons, but this is optional.

Expression value conversion based on information content

Information content measures the average rate of information from data. The smaller the *p*-value, the greater the amount of information. For the adjusted *p*-value of enrichment of a gene set, the null hypothesis is that there is no significant enrichment. So, the smaller the adjusted *p*-value, the more significant the enrichment of the gene set (rather than stochastic). Therefore, here the information content was used as a measure of the level of expression of a functional group.

$$GS_{i,j} = -\log\left(adj - p_{i,j}\right) \quad (2)$$

Here, i is the ith gene set, j is the jth cell, and $adj - p_{i,j}$ represents the adjusted *p*-value in the ith gene set in the jth cell.

Algorithm optimization

Fisher's exact test is a time-consuming process. For single-cell data, a statistical test would be required for each function of each cell. Therefore, when the number of cells is large, the computation time would be untenable. Therefore, the algorithm was optimized with the addition of multi-core computing.

Optimization of the algorithm can be illustrated using the symbols in 2.4.2. First, because N and K were invariable for all cells, these values were stored in memory to avoid recalculation every time. Second, the gene expression matrix was transformed into 0,1 matrix A, where 1 represents the expression of gene i in cell j and 0 represents the non-expression of gene i in cell j. The gene set was also transformed into 0,1 matrix B, where 1 represents the presence of gene i in set s and 0 represents the absence of gene i in set s. The element (k) in the product $A \times B^T$ of the two matrices represents the number of genes expressed by cell j in set s (***Fig. 2***).

Cluster and differentially expressed gene detection based on FEM and integration of GEM and FEM

Analysis tools for snRNA-SEQ data

There are many integrated data analysis software packages and platforms, such as Seurat (Butler et al., 2018), Scater (McCarthy et al., 2017), and Scanpy (Wolf, Angerer & Theis, 2018). Seurat provides integrated environments (including sample and feature selection, data standardization, dimensionality reduction, clustering, and visualization) to explore massive scRNA-SEQ datasets(Luecken & Theis, 2019). This study used Seurat for normalization, dimensionality reduction, clustering, and visualization.

Dimensionality reduction

After feature selection, the dimensions of single-cell expression matrices can be further reduced by dedicated dimensionality-reduction algorithms. These algorithms, such as principal component analysis (PCA), embed the expression matrix into a low-dimensional space, which is designed to capture the underlying structure in the data in as few dimensions as possible(Luecken & Theis, 2019) (Eraslan et al., 2019). In the present study, the data were converted into a linear combination of the first N principal components by the PCA algorithm. The value corresponding to the 'elbow' point was taken as the value of N.

Clustering

A core step in the analysis of scRNA-SEQ transcriptome profiles is to cluster the single cells. This can reveal cell subtypes and infer cell lineages based on the relationships among cells. Several software packages support the cluster analysis of scRNA-SEQ data (Petegrosso, Li & Kuang, 2019); here, Seurat was used to clustering, which is based on a graphical approach. The parameter used in the cluster function was set to 0.4–1.2 according to the circumstances.

Differential expression between clusters

Differential expression analysis is very useful for finding the significant DEG between distinct subpopulations or groups of cells (Petegrosso, Li & Kuang, 2019). Seurat was used to finding subsets of functions that exhibited high variation between clusters.

GEM and FEM fusion analysis

The main problem in FEM is that it only considers the presence or absence of gene expression without considering the expression value of the gene. Therefore, FEM cannot replace GEM-based methods in cell classification and type identification. In the present study, data from the

GEM and FEM were used for fusion analysis (***Fig. 1***). To combine the GEM and FEM results, the multi-modal data analysis module of Seurat was used for polymerization analysis (Stuart et al., 2019) (Stuart & Satija, 2019). The feature number selection, scaling ratio, PCA, and clustering parameter selection were appropriately adjusted according to circumstances, following the Seurat instructions.

Results

FEM can separate different cell types

Since the Reactome pathway gene sets have only 5741 unique genes, a large amount of gene information will be lost for the FEM algorithm. GO gene sets have 15578 unique genes. Therefore, GO gene sets-based FEM was used instead of GEM for cluster analysis. The Reactome pathway gene sets were used in the analysis of GC-DEF.

To verify whether the GO-based FEM algorithm can separate cells of different types, three data sets (PBMC, liver, pancreas) were used to replace GEM for dimensionality reduction and clustering. Since the number of clusters is artificially adjusted by parameters, to better distinguish different cell types, the number of clusters is set to be greater than or equal to the number of actual cell types so that a cluster contains only one main cell type. The results show that on the liver and pancreas data sets, the FEM method can distinguish different types of cells (***Fig. 3***)

Some cells in the PBMC data set have different subtypes, such as FCGR3A+ Mono cells and CD14+ Mono cells, Naïve CD4 T cells, and memory CD4 T cells (***Fig. 4***). There is a large overlap between the different subtypes of these two types of cells in the GO-based FEM cluster. This means that the GO-based FEM method can reflect the "functional similarity" between cells If the two groups of cells are far apart on GEM, but are close or partially overlapped on FEM, it indicates that they are different subtypes of the same cell type, or the two groups of cells may perform similar functions.

FC-DEF can directly detect the functional differences between clusters

In the PBMC dataset, official data pre-processing included the removal of cells with excessive mitochondrial genes (> 5%, quality control) and those with too many (> 2,500) or too few (< 200) features (Bittersohl & Steimer, 2016). Therefore, some cells may have been filtered out in the data pre-processing stage. The present method did not filter out any cells. Indeed, the filtered cells were found to be located above the NC cells. Using the GC-DEF method, this cell group was found to have a highly expressed cell proliferation-related pathway. Thus, these cells were identified as proliferative (***Fig. 5***).

The most significantly expressed pathway, “hemostasis,” was located in the platelet clusters. The “innate immune system” pathway was significantly expressed in the monocyte, DC cell, NK cell, and platelet cell populations, which was consistent with literature results(NORMAN, 1995). The top five highly expressed functions were consistent with the cell type character. All other results of FC-DEF and GC-DEF are provided in Table s1.

Table 2 shows that most of the corresponding cells were closely related to their corresponding top functions, such as the high expression of "Reactome regulation of beta cell development" in beta cells and the high expression of "Reactome gluconeogenesis" in mature stem cells. All other results are shown in Table s1.

GO-based GC-DEF results was consistent with the pathway-based methods and literature(NORMAN, 1995). All other results are presented in Table 3 and Table s1, s2, s3 and s4.

Validation with an immune dataset

To test whether the proposed method could detect sets of genes that had been identified as up-regulated or down-regulated by traditional methods, the Immunologic Signatures Collection (ImmuneSigDB) (Liberzon et al., 2011) was employed as a validation dataset.

The ImmuneSigDB is composed of gene sets that represent cell types, states, and perturbations within the immune system(Godec et al., 2016). Figure 6 shows that in the bulk RNA-SEQ dataset, the cell types from the up-regulated expression marker gene set were also highly expressed using the method proposed in this study. This demonstrated the efficacy of the proposed method for detecting cell-type-specific gene sets.

Discussion

We prove that FEM can be used for cell clustering. It also can replace or merge the GEM method for downstream differential expression analysis to find cell type-specific functions.

Sometimes, evaluate the impact of some biological effects on the cell clusters is necessary, such as the impact of cell cycles on cell-type clustering results. However, there is no uniform standard for how to achieve this. Because the proposed method directly converts the expression of genes in cells to the expressions of functions, cells can be screened according to their FEM score at any stage of processing.

It should be noted that FEM only considers the presence and absence of gene expression, without considering the influences of gene-expression values, so the proposed method and gene-expression–based methods are complementary rather than alternatives.

Conclusions

Usually, the final step in the analysis of gene expression data is to interpret the biological significance of the genes based on GSEA. However, if each of the first n genes obtained by differential expression analysis represents a function that does not overlap with each other, then the enrichment analysis will fail. On the other hand, if most genes of a function are expressed in a cell and are screened out in the process of gene selection, this function will also be missed. Based on the characteristics of single-cell data, GSEA at the single-cell level effectively avoids the above problems. The results of the present study showed that direct enrichment analysis at the single-cell level is feasible and powerful.

Acknowledgements

We thank Southeast University State Key Laboratory of Bioelectronics for provide the computing resource for our research.

References

Bittersohl H, Steimer W. 2016. Intracellular Concentrations of Immunosuppressants. In: *Personalized Immunosuppression in Transplantation: Role of Biomarker Monitoring and Therapeutic Drug Monitoring*. Elsevier Inc., 199–226. DOI: 10.1016/B978-0-12-800885-0.00009-6.

Blasi T, Buettner F, Strasser MK, Marr C, Theis FJ. 2017. CgCorrect: A method to correct for confounding cell-cell variation due to cell growth in single-cell transcriptomics. *Physical Biology*. DOI: 10.1088/1478-3975/aa609a.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. DOI: 10.1038/nbt.4096.

Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D, Belicova L, Bickle M, Barsacchi R, Okuda R, Yoshizawa E, Kimura M, Ayabe H, Taniguchi H, Takebe T, Treutlein B. 2017. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546. DOI: 10.1038/nature22796.

Chen G, Ning B, Shi T. 2019. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics* 10:1–13. DOI: 10.3389/fgene.2019.00317.

Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* 10:1–14. DOI: 10.1038/s41467-018-07931-2.

Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. 2018. Single sample scoring of molecular phenotypes. *BMC Bioinformatics* 19:1–10. DOI: 10.1186/s12859-018-2435-4.

Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics* 17:175–188. DOI: 10.1038/nrg.2015.16.

Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, Mesirov JP, Haining WN. 2016. Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity* 44:194–206. DOI: 10.1016/j.immuni.2015.12.006.

Griffiths JA, Scialdone A, Marioni JC. 2018. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*. DOI: 10.15252/msb.20178046.

Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*. DOI: 10.1093/bioinformatics/btv325.

Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14:1–21. DOI: 10.1186/1471-2105-14-7.

Hicks SC, Townes FW, Teng M, Irizarry RA. 2018a. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*. DOI: 10.1093/biostatistics/kxx053.

Hicks SC, Townes FW, Teng M, Irizarry RA. 2018b. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19:562–578. DOI: 10.1093/biostatistics/kxx053.

Hwang B, Lee JH, Bang D. 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* 50. DOI: 10.1038/s12276-018-0071-8.

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*. DOI: 10.1186/s13059-016-0888-1.

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* 20:273–282. DOI: 10.1038/s41576-018-0088-9.

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. 2015. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*. DOI: 10.1016/j.molcel.2015.04.005.

Kulkarni A, Anderson AG, Merullo DP, Konopka G. 2019. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology* 58:129–136. DOI: 10.1016/j.copbio.2019.03.001.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. DOI: 10.1038/nrg2825.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740. DOI: 10.1093/bioinformatics/btr260.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* 15:e8746. DOI: 10.15252/msb.20188746.

McCarthy DJ, Campbell KR, Lun ATL, Wills QF. 2017. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*. DOI: 10.1093/bioinformatics/btw777.

McDavid A, Finak G, Gottardo R. 2016. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nature Biotechnology*. DOI: 10.1038/nbt.3498.

McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3. DOI: 10.21105/joss.00861.

Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, van Gurp L, Engelse MA, Carlotti F, de Koning EJP, van Oudenaarden A. 2016. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* 3. DOI: 10.1016/j.cels.2016.09.002.

NORMAN P. 1995. Immunobiology: The immune system in health and disease. *Journal of Allergy and Clinical Immunology*. DOI: 10.1016/s0091-6749(95)70025-0.

Petegrosso R, Li Z, Kuang R. 2019. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics* 00:1–15. DOI: 10.1093/bib/bbz063.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* 177:1888-1902.e21. DOI: 10.1016/j.cell.2019.05.031.

Stuart T, Satija R. 2019. Integrative single-cell analysis. *Nature Reviews Genetics* 20:257–272. DOI: 10.1038/s41576-019-0093-7.

Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park JE, Stephenson E, Polański K, Goncalves A, Gardner L, Holmqvist S, Henriksson J, Zou A, Sharkey AM, Millar B, Innes B, Wood L, Wilbrey-Clark A, Payne RP, Ivarsson MA, Lisgo S, Filby A, Rowitch DH, Bulmer JN, Wright GJ, Stubbington MJT, Haniffa M, Moffett A,

Teichmann SA. 2018. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*. DOI: 10.1038/s41586-018-0698-6.
Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*. DOI: 10.1186/s13059-017-1382-0.

Table 1(on next page)

Functional gene set

1

Function name	Number of gene sets	details
C2: Reactome gene sets	1,499	Gene sets derived from the Reactome pathway database.
C5: GO gene sets	9,996	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: BP, CC, and MF.
C7: immunologic signature gene sets	4,872	Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology.

2

3

4

Table 2(on next page)

Top five pathways for mature hepatocyte cells (liver dataset) and beta cells (pancreas dataset)

1

Function	cluster	Adjusted p-value
reactome-regulation-of-gene-expression-in-beta-cells	beta	7.47E-48
reactome-regulation-of-beta-cell-development	beta	1.02E-25
reactome-activation-of-nmda-receptors-and-postsynaptic-events	beta	
reactome-negative-regulation-of-tcf-dependent-signaling-by-dvl-interacting-proteins	beta	2.05E-11
reactome-synthesis-of-pips-at-the-early-endosome-membrane	beta	1.76E-06
reactome-gluconeogenesis	mature hepatocyte	1.75E-11
reactome-signaling-by-bmp	mature hepatocyte	4.50E-09
reactome-apoptotic-cleavage-of-cell-adhesion-proteins	mature hepatocyte	9.00E-09
reactome-transport-of-nucleosides-and-free-purine-and-pyrimidine-bases-across-the-plasma-membrane	mature hepatocyte	1.51E-08
reactome-bbsome-mediated-cargo-targeting-to-cilium	mature hepatocyte	1.76E-08

2
3
4

Table 3(on next page)

Results of the top five GO-based FEM methods for each cluster

Function	cluster	Adjusted p-value
go-mhc-class-ii-receptor-activity	B	1.31E-299
go-mhc-class-ii-protein-complex	B	1.94E-207
go-clathrin-coated-endocytic-vesicle-membrane	B	8.60E-183
go-clathrin-coated-endocytic-vesicle	B	1.40E-174
go-mhc-protein-complex-assembly	B	2.39E-173
go-collagen-containing-extracellular-matrix	CD14+_Mono	3.17E-235
go-rage-receptor-binding	CD14+_Mono	1.42E-226
go-chemokine-production	CD14+_Mono	1.61E-211
go-defense-response-to-bacterium	CD14+_Mono	9.86E-194
go-neutrophil-migration	CD14+_Mono	2.36E-187
go-cytolysis	CD8_T	2.61E-64
go-t-cell-receptor-complex	CD8_T	7.62E-53
go-negative-regulation-by-host-of-viral-transcription	CD8_T	5.71E-39
go-regulation-of-cell-cell-adhesion-mediated-by-integrin	CD8_T	1.50E-36
go-t-cell-receptor-binding	CD8_T	4.19E-29
go-ige-binding	DC	4.78E-241
go-t-cell-activation-via-t-cell-receptor-contact-with-antigen-bound-to-mhc-molecule-on-antigen-presenting-cell	DC	6.80E-23
go-mhc-class-ii-receptor-activity	DC	1.48E-14
go-hydrolase-activity-acting-on-ester-bonds	DC	0.000373
go-lipid-metabolic-process	DC	0.000529
go-igg-binding	FCGR3A+_Mono	9.13E-143
go-negative-regulation-of-leukocyte-proliferation	FCGR3A+_Mono	4.71E-60
go-regulation-of-mast-cell-activation	FCGR3A+_Mono	7.62E-59
go-regulation-of-mast-cell-activation-involved-in-immune-response	FCGR3A+_Mono	1.83E-58
go-dendritic-cell-differentiation	FCGR3A+_Mono	3.00E-58
go-positive-t-cell-selection	Memory_CD4_T	5.51E-40
go-positive-thymic-t-cell-selection	Memory_CD4_T	1.80E-37
go-t-cell-receptor-binding	Memory_CD4_T	1.89E-37
go-alpha-beta-t-cell-receptor-complex	Memory_CD4_T	7.89E-35
go-positive-regulation-of-t-cell-receptor-signaling-pathway	Memory_CD4_T	3.62E-33
go-t-cell-differentiation-in-thymus	Naive_CD4_T	5.58E-123
go-thymic-t-cell-selection	Naive_CD4_T	4.23E-87
go-positive-regulation-of-t-cell-receptor-signaling-pathway	Naive_CD4_T	2.06E-58
go-t-cell-receptor-complex	Naive_CD4_T	2.18E-57
go-negative-t-cell-selection	Naive_CD4_T	1.51E-39
go-granzyme-mediated-apoptotic-signaling-pathway	NK	2.10E-190
go-cytolytic-granule	NK	1.03E-150
go-positive-regulation-of-natural-killer-cell-chemotaxis	NK	6.27E-128
go-cytolysis	NK	6.91E-97
go-ccr5-chemokine-receptor-binding	NK	9.03E-58
go-platelet-alpha-granule-membrane	Platelet	6.75E-158
go-platelet-alpha-granule	Platelet	4.94E-06
go-platelet-degranulation	Platelet	5.90E-06
go-platelet-alpha-granule-lumen	Platelet	6.23E-06
go-contractile-fiber	Platelet	8.58E-06

Figure 1

Workflow of the proposed method.

(A) Flowchart of the FEM algorithm, which was used to calculate Fisher's exact test for each cell and each gene set, following which the calculated p-value matrix was converted into the FEM through information content. The efficiency of the calculation method was improved by matrix operation and multi-core parallel processing (see the method section for details). (B) Cluster differential expression analysis based on FEM and integration of FEM and GEM. (C) FEM was used for dimension reduction, clustering, and differential expression function analysis (FC-DEF). (D) Due to the loss of gene expression information in FEM, the dimensionality and clustering were first reduced based on GEM, and differential expression function analysis among clusters in the GEM cluster (GC-DEF) was performed.

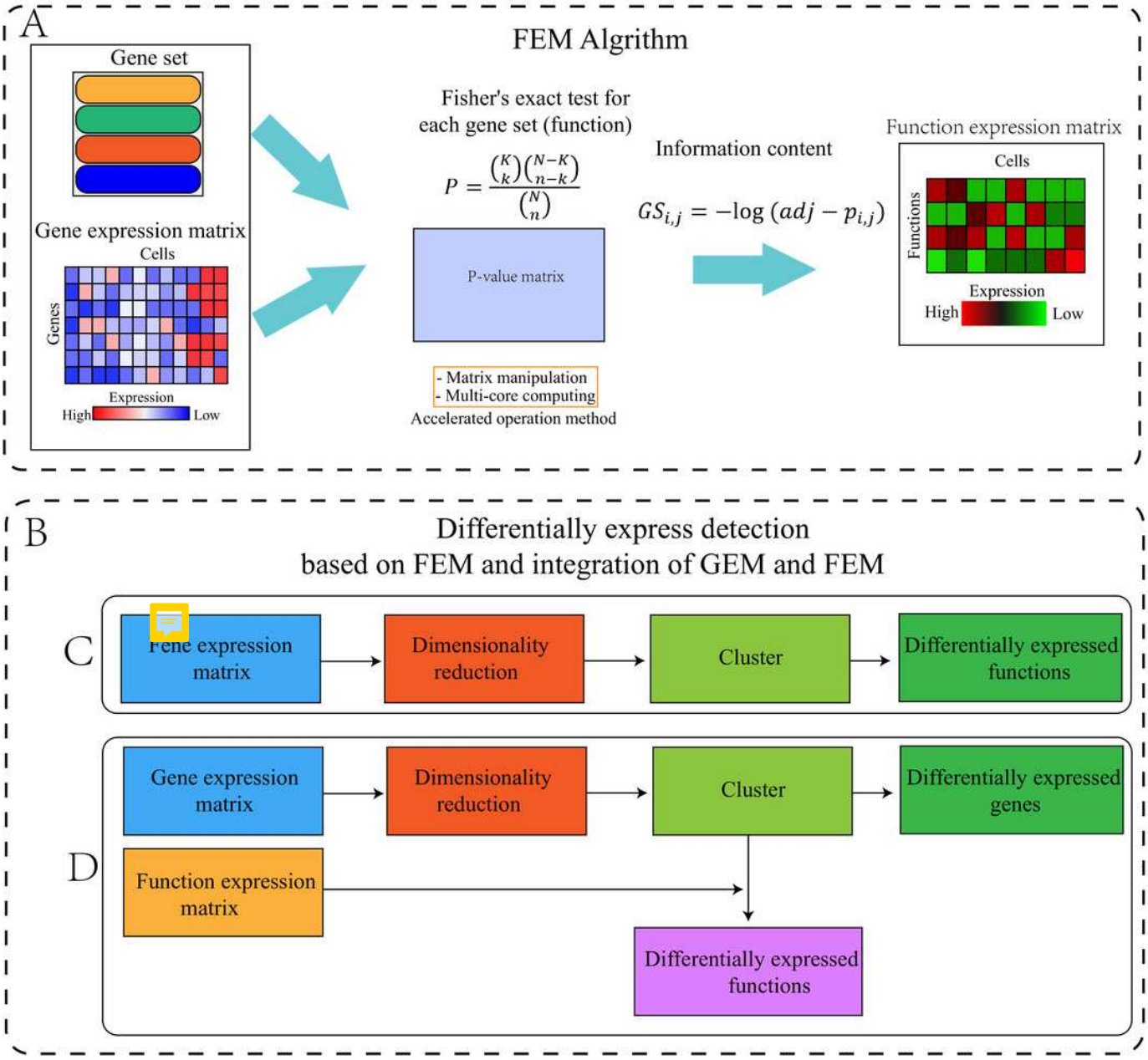

A
FEM Algrithm
Gene set
Gene expression matrix
Cells
Genes
Expression
High
Low
Fisher's exact test for each gene set (function)
$P = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
P-value matrix
- Matrix manipulation
- Multi-core computing
Accelerated operation method
Information content
$GS_{i,j} = -\log(adj - p_{i,j})$
Function expression matrix
Cells
Functions
Expression
High
Low
B
Differentially express detection
based on FEM and integration of GEM and FEM
C
Gene expression matrix
Dimensionality reduction
Cluster
Differentially expressed functions
D
Gene expression matrix
Dimensionality reduction
Cluster
Differentially expressed genes
Function expression matrix
Differentially expressed functions

Figure 2

Algorithm optimization based on matrix multiplication and multi-core operation.

(A) Convert the GEM and the gene sets into a 0, 1 matrix. The result of multiplication of the two matrices representing the number of genes expressed in cell *j* and set *s*. (B) The multi-core parallel computing method established two queues. The data queue was used to store the data needed for calculation and the result queue stored the calculated results. Each set of the two queues uniquely identified the cell and the function to which it belonged. The calculation process adopted multi-process operation.

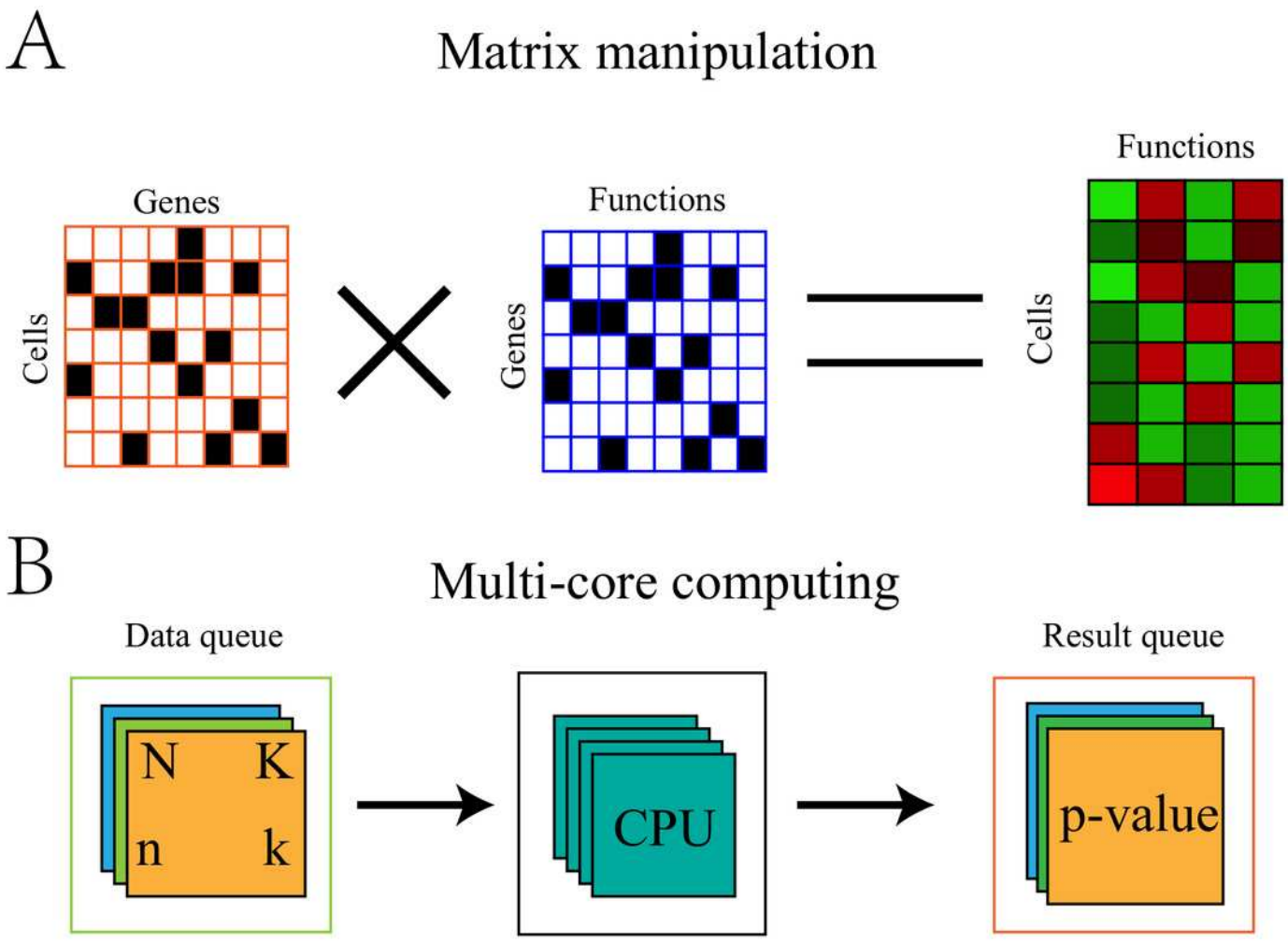
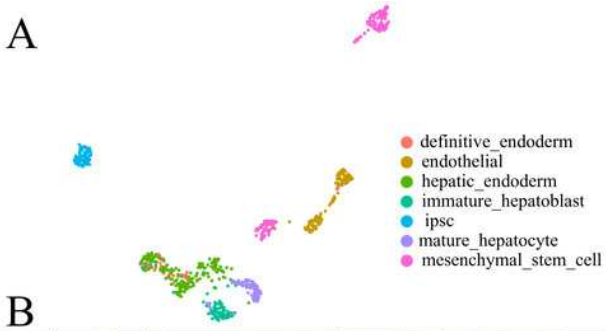


Figure 3

The clustering results of the GO-based FEM algorithm.

(A) GO-based FEM cluster for each cell type in liver data set. (B) The overlap between GO-based FEM cluster and ground truth cell label in liver data set. (C) GO-based FEM cluster for each cell type in pancreas data set. (D) The overlap between GO-based FEM cluster and ground truth cell label in pancreas data set.



B

cell type / cluster	hepatic endoderm	endothelial	mesenchymal stem cell	immature hepatoblast	mature hepatocyte	definitive endoderm	ipsc
5	0	0	0	0	73	0	0
2	0	0	0	0	0	0	80
1	0	0	88	0	0	0	0
9	0	53	0	0	0	0	0
8	0	0	63	0	0	3	0
4	54	0	0	11	8	0	0
7	1	0	0	0	0	67	0
6	0	0	0	69	0	0	0
0	112	0	0	1	0	0	0
3	0	74	0	0	0	0	0
10	0	0	20	0	0	0	0

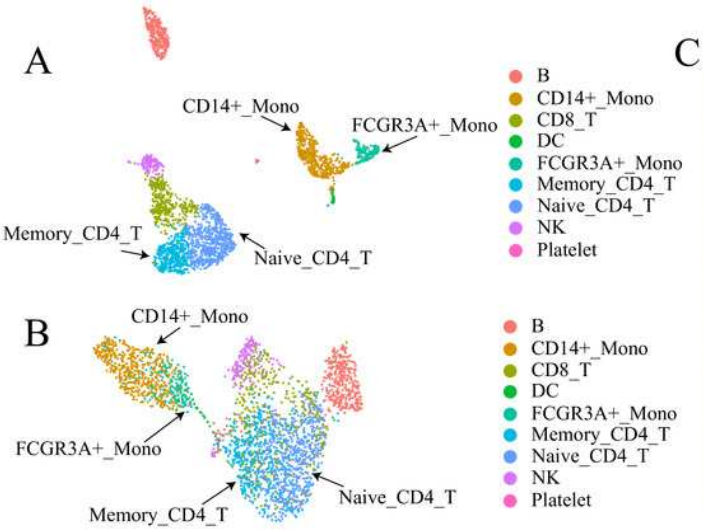
D

cell type / cluster	acinar	mesench-ymal	unclear	delta	alpha	gamma	beta	endothe-lial	ductal	epsilon
7	0	0	0	1	0	0	148	0	0	0
6	0	0	0	185	0	0	0	0	0	0
8	0	0	0	0	0	99	6	0	0	2
4	1	1	4	0	6	0	5	0	212	0
5	0	0	0	0	219	0	0	0	0	1
12	0	0	0	0	0	0	1	0	15	0
0	0	0	0	0	322	0	1	0	0	0
1	0	0	0	3	16	0	239	0	0	0
3	217	0	0	0	1	2	3	0	8	0
11	0	0	0	0	0	0	0	20	0	0
10	0	0	0	4	2	0	42	0	4	0
2	0	0	0	0	246	0	1	0	0	0
9	1	79	0	0	0	0	2	1	6	0

Figure 4

Comparison of cluster results based on GEM and FEM.

(A) Clustering results based on GEM. (B) Clustering results based on FEM. (C) The overlap between GO-based FEM cluster and ground truth cell label in PBMC dataset.



cluster \ cell type	CD8 T	B	FCGR3A+ Mono	CD14+ Mono	Memory CD4_T	Naive CD4_T	Platelet	NK	DC
3	0	0	18	317	0	0	0	0	1
4	320	1	0	0	1	2	0	0	0
6	12	74	0	0	6	127	0	3	0
5	0	115	0	0	3	0	0	135	0
8	0	0	0	0	0	0	11	0	0
7	8	3	0	0	12	3	0	12	5
2	0	0	143	175	0	0	1	0	29
0	2	77	0	0	116	418	0	0	0
1	8	114	0	0	289	134	1	4	0

Figure 5

An example of GC-DEF functional analysis.

(A) Official clustering results. (B) Results of GEM without cell filtering show a small cell group above the NK cell cluster. (C) By directly displaying the expression value of specific pathways in all cells, this cell group was found to have high expression of the cell proliferation-related pathway.

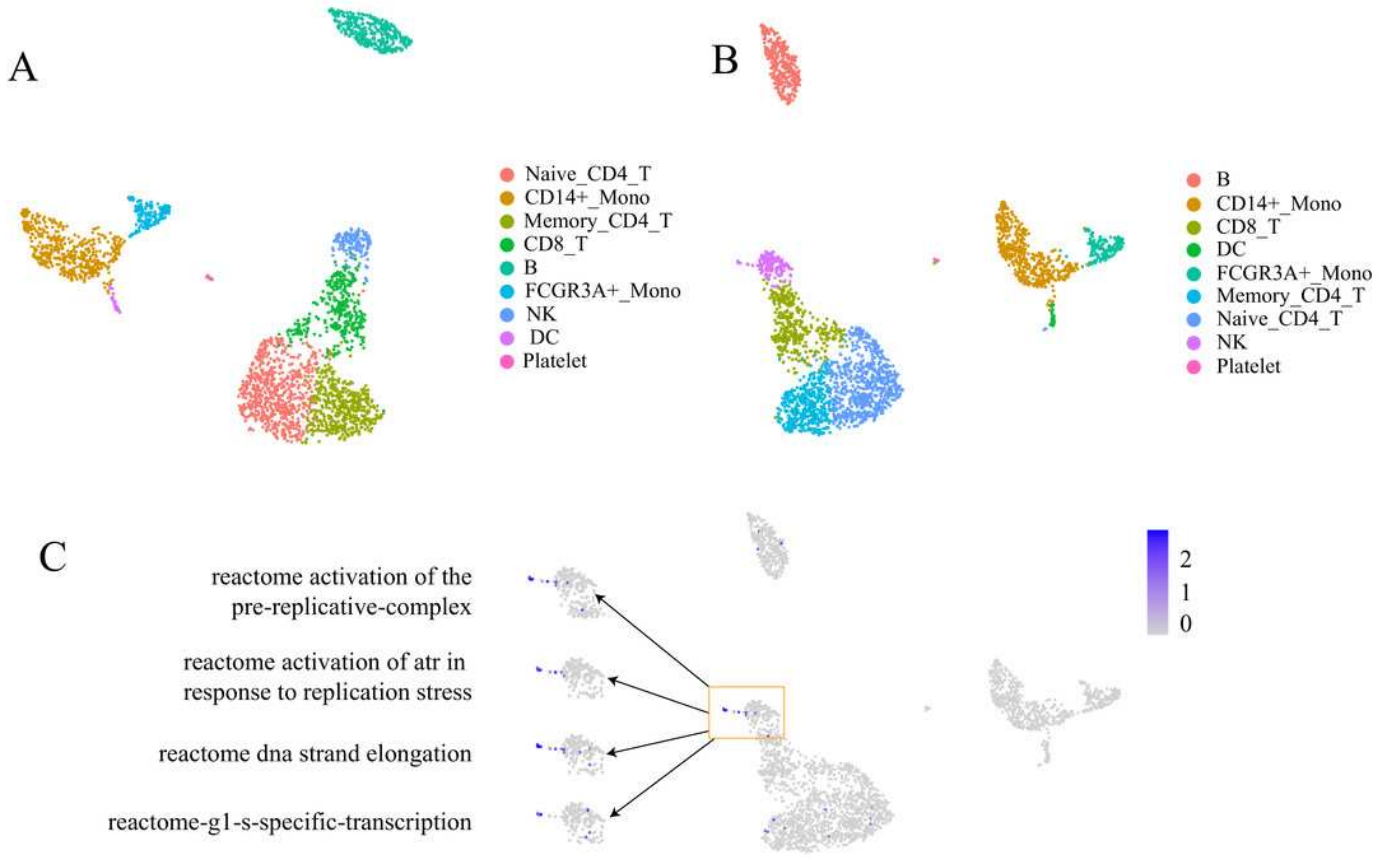


Figure 6

Validation results of Immunologic Signatures Collection gene sets.

Each of these gene sets represents a set of all highly expressed genes of one (or several) cell types relative to another (or several) cell types.