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], **Guo Zhuojun** [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. Gene Set Enrichment (GSE) analysis is one of the main tools for function mining that has been widely used. However, every gene expressed in a cell is valuable information for GSE for single-cell RNA sequencing (scRNA-SEQ) data and not should be discarded. **Methods.** We developed the functional expression matrix (FEM) algorithm to utilize the information from all expressed genes. The algorithm converts the gene expression matrix (GEM) into a FEM. The FEM algorithm can provide insight on the biological significance of a single cell. It can also integrate with GEM for downstream analysis. **Results.** We found that FEM performed well with cell clustering and cell-type specific function annotation in three datasets (peripheral blood mononuclear cells, human liver, and human pancreas).

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

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

Abstract

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

Methods. We developed the functional expression matrix (FEM) algorithm to utilize the information from all expressed genes. The algorithm converts the gene expression matrix (GEM) into a FEM. The FEM algorithm can provide insight on the biological significance of a single cell. It can also integrate with GEM for downstream analysis.

Results. We found that FEM performed well with cell clustering and cell-type specific function annotation in three datasets (peripheral blood mononuclear cells, human liver, and human pancreas).

Introduction

RNA sequencing (RNA-seq) has been widely used as an alternative to the microarray platform over the past 10 years. It has provided valuable insight into complex biological mechanisms, ranging from cancer genomics to diverse microbial communities. Single-cell RNA-seq (scRNA-seq) can facilitate new and potentially unexpected biological discoveries when compared with traditional bulk methods that profile batches of pooled cells. For example, it can provide information on complex and rare cell populations and 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).

Cells are the fundamental unit in biology. Biologists have known for centuries that multicellular organisms are characterized by a plethora of distinct cell types (Kiselev, Andrews & Hemberg, 2019).

It is difficult to distinguish cell types from the genome sequence since all cells of an individual share the same set of genetic material. Their differences may be characterized by transcriptome similarity as defined by unsupervised clustering (Kiselev, Andrews & Hemberg, 2019).

Bulk transcriptome studies using samples containing multiple cells cannot distinguish between the cells within the sample. Single-cell transcriptomics can address this issue (Kulkarni et al., 2019). High-throughput single-cell transcriptomics have provided unprecedented insights into the cellular diversity of tissues across diverse organisms. scRNA-seq is a promising approach in the study of the transcriptomes of organism's individual cells.

The data processing pipeline for scRNA-seq is based on bulk RNA-seq. The basic workflow includes 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). Many genes are expressed through RNA-seq data, which is made up of pooled samples of multiple cells. However, it remains unclear whether these genes are co-expressed or just the result of

81 mixing. Therefore, a portion of the particularly well-characterized genes must be screened for
82 downstream analysis.
83   This type of scRNA-seq data analysis has specific shortcomings. First, the characteristics of
84 single-cell data are not fully utilized. scRNA-seq data are a direct reflection of the physiological
85 state of a cell, but the aforementioned method is an analytical process based on multiple cells or
86 cell clusters. Second, it cannot effectively capture all meaningful functional groups since many
87 genes are filtered but may still play important roles in transient cell states. It is possible to miss
88 meaningful biological functions since functional enrichment analysis is based on all of the cell's
89 expressed genes, including genes that were previously filtered out. Third, the results of
90 downstream functional analysis and upstream clustering are not integrated and visualized. To
91 overcome these limitations, this study proposes a method based on the functional expression
92 matrix (FEM) which converts each cell's gene expression into an enriched gene set representing
93 a specific function.
94

95 **Materials & Methods**
96 **Motivation for this approach**
97 Functional Gene Set Enrichment Analysis (GSEA) is often the last step of expression data
98 analysis. A gene set usually represents a biological function, and the function will be used to
99 refer to a gene set in subsequent analyses. Many R packages and some online sites, such as
100 DAVID, provide enrichment analysis tools. The software takes a set of genes given by the user
101 as input and returns functions (pathways) that are significantly enriched. Enrichment analysis
102 methods for individual samples have also been proposed (Foroutan et al., 2018; Hänzelmann,
103 Castelo & Guinney, 2013).
104   Zhang et al. (2020) summarized the widely used GSEA algorithms. We divided their methods
105 into three categories: methods based on dimension reduction, methods based on single samples,
106 and methods based on phenotype or cluster.
107   Methods based on dimension reduction include Pagoda2 (Fan et al., 2016) and PLAGE
108 (Tomfohr, Lu & Kepler, 2005). These methods convert the gene expression matrix into FEM by
109 performing dimension reduction (SVD for PLAGE and PCA for Pagoda2) on each gene set
110 (pathway). Dimension reduction uses all samples and is not entirely based on a single cell.
111   Methods based on phenotype or cluster include Vision (DeTomaso et al., 2019) and z-score
112 (Lee et al., 2008). Vision methods clusters GEM into clusters by KNN, and GSEA is based on
113 these clusters. The Z-score method depends on phenotypes, which define each pathway activity
114 score based on the top gene of t-test score between two conditions. These gradually join the
115 genes in pathways to maximize the activity score and are not based on a single sample.
116   Methods based on the single sample include AUCell (Aibar et al., 2017), ssGSEA (Barbie et
117 al., 2009), and GSVA (Hänzelmann, Castelo & Guinney, 2013), which are all non-parametric
118 test statistical methods based on gene rank. Data sparsity is the main problem when analysing
119 data from a single cell. Therefore, the level of gene expression is not the primary consideration

120   for a pathway, but rather, whether the gene is expressed. A method based on Fisher's exact test
121   or the chi-square test may have better results than the no-parameter test method based on ranking.
122   scRNA-seq has low capture efficiency and high dropouts due to the small amount of initial
123   material used. scRNA-seq expression data tend to be sparse (Hicks et al., 2018b; Eraslan et al.,
124   2019). A true zero represents the lack of expression of a gene in a specific cell due to the non-
125   trivial distinction between true- and false-zero counts. Meanwhile, a false zero is a technical
126   deviation. The RNA-seq technique only detects mRNA molecules that are present, therefore, a
127   gene in the scRNA-seq dataset with a non-zero expression value indicates that at least one
128   mRNA molecule is present. These sparse non-zero expression values provided inspiration for the
129   possibility of performing functional GSEA analysis at the single-cell level.
130   Functional enrichment analysis may be performed at the single-cell transcriptome expression
131   level given the characteristics of scRNA-seq data. The algorithm is as follows: first, replace the
132   initial expression matrix with a FEM that allows for the direct exploration of the differences
133   between cells from different biological perspectives. Second, the top variable function of a cell
134   cluster obtained by a gene expression matrix (GEM) can be represented as a cell scatter plot.
135   Third, the coverage of a functional gene set is more important than the expression of individual
136   genes for GSEA since genes often overlap different functional groups, which makes single high-
137   expression genes difficult to explain. For example, the activation of a signalling pathway is the
138   result of interactions between all genes; the high expression of one gene does not mean that the
139   pathway is activated.
140   The current factors affecting the discovery of cell groups are as follows: first, technical
141   covariates must be regressed out before downstream analysis can occur since these factors
142   introduce systematic error and confound the technical and biological variability. This error may
143   lead to systematic differences in gene expression profiles between batches (Leek et al., 2010;
144   Hicks et al., 2018a; Chen, Ning & Shi, 2019). The most prominent technical covariates in single-
145   cell data are *count depth* and *batch.* The FEM method is based on a gene set, which makes it
146   more robust than the GEM. The measured expression levels for the genes related to a certain
147   biological process may fluctuate between batches or single cells due to variations in sequencing
148   processing. That is, detection for one specific high-expression gene in single cell has a certain
149   randomness. If we can consider the expression of a set of genes related to a specific process, we
150   can increase the detection stability for the biological process of single cells, since the number of
151   expressed genes of a specific biological process obey a hypergeometric distribution. Second,
152   some biological effects may affect the results of the cluster algorithms. For example, the cell
153   cycle phase of a given cell within a cell type population may cause separation from the same
154   cells in the clustering analysis. However, correcting for biological covariates is not always
155   helpful in interpreting scRNA-seq data (Kolodziejczyk et al., 2015). These influencing factors
156   often do not have uniform filtering criteria. In some cases, the cell cycle may be part of the
157   study, or there may be a relationship between the cell cycle and other functions (Haghverdi,
158   Buettner & Theis, 2015; Vento-Tormo et al., 2018 ; McDavid, Finak & Gottardo, 2016; Blasi et

159 al., 2017). The FEM method can be used to systematically survey the biological aspect of each
160 cell before downstream analysis. To this end, we developed an scRNA-seq FEM algorithm.

Workflow of the proposed methods

162 The FEM was divided into four steps (Fig. 1). **Step 1:** The standardized scRNA-seq GEM was
163 transformed into a FEM by multi-module gene enrichment analysis (gene ontology, pathway).
164 **Step 2:** The *p*-value represents the significance of enrichment, therefore, the *p*-value obtained by
165 enrichment analysis was converted into information content (FEM). **Step 3:** FEM/GEM was
166 used for data standardization, dimensionality reduction clustering, and UMAP visualization
167 (McInnes et al., 2018). **Step 4:** Differentially expressed genes (DEG) and differentially
168 expressed functions (DEF) analysis was performed based on GEM clustering (see the "GEM and
169 FEM fusion analysis" section for details).

Dataset

171 Peripheral blood mononuclear cells (PBMCs) are populations of immune cells that remain in the
172 less dense upper interface of the Ficoll layer. PBMCs include lymphocytes (T cells, B cells, and
173 NK cells), monocytes, and dendritic cells (DCs). The frequencies of these populations vary
174 across individuals in humans. Lymphocytes are typically in the range of 70–90%, monocytes
175 range from 10–30%, while DCs are only present at 1–2% (NORMAN, 1995). The PBMCs
176 dataset used in this study was downloaded from the 10X Genomics official website
177 (https://www.10xgenomics.com/resources/datasets), which included B cells, NK cells, CD8 T
178 cells, memory CD4 T cells, naïve CD4 T cells, DC, CD14+ monocytes, FCGR3A+ monocytes,
179 and a small number of platelets. This dataset contained 2,700 cells in total.
180   The human pancreas dataset contained 2,126 cells and 10 cell types. This included alpha,
181 ductal, endothelial, delta, acinar, beta, gamma, mesenchymal, and epsilon cells, as well as a
182 small number of unknown cell types (Muraro et al., 2016).
183   The human liver dataset consisted of 777 cells, including seven types of cells: definitive
184 endoderm cells, immediate hepatoblast cells, induced pluripotent stem cells (IPSCs), material
185 hepatocytic cells, hepatic endoderm cells, endothelial cells, and mesenchymal stem cells (Camp
186 et al., 2017).

FEM algorithm

Selected functional groups and their profiles

189 Three functional gene sets were selected from the Msigdb (v6.2) database for enrichment
190 analysis: the Reactome pathways, gene ontology (GO), and immunologic signature gene sets
191 (Liberzon et al., 2011) (Table 1).

Validation gene set

193 ImmuneSigDB (Godec et al., 2016) contained a gene set composed of differentially expressed
194 genes from human and mouse immune-related cells. These were collected from the literature
195 with corresponding expression data in the Gene Expression Omnibus (GEO) database. All data
196 were manually reviewed and standardized using the same method. The differentially expressed
197 FDR value was less than 0.02, and the genes were sorted using the mutual information algorithm.
198 No more than 300 genes were selected for each gene set.

Gene-functional group conversion

199
200 The cells' non-zero–expressed genes (i.e., genes in which the read count or normalized expression values,
201 such as RPKM and FPKM, were not zero) in the GEM were extracted first. In the second step, an
202 enrichment analysis score for each cell was calculated. The enrichment analysis method was
203 based on Fisher's exact test using the Python SciPy package. The Fisher exact test is a statistical
204 test based on a hypergeometric distribution that is used to determine if there are non-random
205 associations between two categorical variables, or to test whether the theoretical value is
206 consistent with the actual value.

$$P = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1}$$

207 Here, N represents the total number of background genes, which is defined as all non-
208 redundant genes in all gene sets taken from a species' function database, such as Reactome
209 pathway or GO database. K represents the number of genes in a particular gene set, n represents
210 the number of non-zero genes in a single cell, and k represents the number of genes present in
211 both K and n. The Bonferroni correction was used to counteract the problem of multiple
212 comparisons, but this is optional.

Expression value conversion based on information content

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

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

219
220 Here, i is the ith gene set, j is the jth cell, and $adj - p_{i,j}$ represents the adjusted p-value in the
221 ith gene set in the jth cell.

222

Algorithm optimization

224 Fisher's exact test is a time-consuming process. For single-cell data, a statistical test would be
225 required for each function of each cell. Therefore, when the number of cells is large, the
226 computation time would be untenable. Therefore, the algorithm was optimized with the addition
227 of multi-core computing.
228 Optimization of the algorithm can be illustrated using the symbols in section "Gene-functional
229 group conversion". First, because N and K were invariable for all cells, these values were stored
230 in memory to avoid recalculation with each replicate. Second, the gene expression matrix was
231 transformed into 0,1 matrix A, where 1 represents the expression of gene i in cell j and 0
232 represents the non-expression of gene i in cell j. The gene set was also transformed into 0,1

233   matrix B, where 1 represents the presence of gene i in set s and 0 represents the absence of gene
234   i in set s. The element (k) in the product $A \times B^T$ of the two matrices represents the number of
235   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 by Seurat

Analysis tools for scRNA-seq data

239   A number of integrated data analysis software packages and platforms exist, including Seurat
240   (Butler et al., 2018), Scater (McCarthy et al., 2017), and Scanpy (Wolf, Angerer & Theis, 2018).
241   Seurat provides integrated environments (including sample and feature selection, data
242   standardization, dimensionality reduction, clustering, and visualization) to explore massive
243   scRNA-seq datasets (Luecken & Theis, 2019). Seurat was used in this study for normalization,
244   dimensionality reduction, clustering, and visualization.

Dimensionality reduction

246   The dimensions of single-cell expression matrices were further reduced after feature selection by
247   dedicated dimensionality-reduction algorithms. These algorithms, such as principal component
248   analysis (PCA), embed the expression matrix into a low-dimensional space, which is designed to
249   capture the underlying structure in the data in as few dimensions as possible (Luecken & Theis,
250   2019; Eraslan et al., 2019). Our data were converted into a linear combination of the first N
251   principal components by the PCA algorithm. The value corresponding to the 'elbow' point was
252   taken as the value of N.

Clustering

254   Single cells were clustered during the analysis of scRNA-seq transcriptome profiles. This may
255   reveal cell subtypes and infer cell lineages based on the relationships among cells. Several
256   software packages support the cluster analysis of scRNA-seq data (Petegrosso, Li & Kuang,
257   2019). Seurat was used for clustering based on a graphical approach. The cluster function was set
258   to 0.4–1.2 according to the data. All parameters are show in Table S5

Differential expression between clusters

260   Differential expression analysis is helpful for finding the significant DEG/DEF between distinct
261   subpopulations or groups of cells (Petegrosso, Li & Kuang, 2019). Seurat was used to determine
262   the subsets of functions that exhibited a high variation between clusters.

GEM and FEM fusion analysis

264   FEM is limited in that it only considers the presence or absence of gene expression without
265   considering the expression value of the gene. Therefore, FEM cannot replace GEM-based
266   methods in cell classification and type identification. In the present study, data from the GEM
267   and FEM were used for fusion analysis (Fig. 1). The multi-modal data analysis module of Seurat
268   was used for polymerization analysis to combine the GEM and FEM results (Stuart et al., 2019;
269   Stuart & Satija, 2019). GEM was used for clustering because it contains information about gene
270   expression values and can better distinguish the differences between cells. FEM reduces all of
271   the genes of the same function to one dimension. If two cells express different genes of the same
272   function, they will be clustered together in FEM, but they may be separated in GEM since they

273   are different genes. We also tested the Differential Expression Functions between Clusters based
274   on GEM (GC-DEF) and the Differentially Expressed Functions between Clusters based on FEM
275   (FC-DEF).
276   The feature number selection, scaling ratio, PCA, and clustering parameter selection were
277   appropriately adjusted according to circumstances (Table S5), following the Seurat instructions.
278   In brief, we use the VlnPlot visualization function in the Seurat package to select the feature
279   number and scaling ratio value for subsequent analysis. The feature number and the scaling ratio
280   are respectively taken as the maximum value of ordinate that can include all the points after
281   excluding individual excessively high outliers. We can select the number of principal
282   components by observing the corresponding abscissa value at the "elbow" point of the Elbow
283   plot by Seurat. If the number of cell types is known, the clustering parameters were set so that
284   the number of clusters was greater than or equal to the number of actual cell types. If the number
285   of cell types is unknown, we can adjust the number of clusters by setting the resolution
286   parameter of FindClusters function in Seurat. Like most clustering algorithms, there is no
287   uniform standard to define a specific number of clusters. Users can try multiple parameters to get
288   a more appropriate number of clusters. All parameters in this paper are detailed in Table S5.

289   **Evaluation of FEM clustering results**

290   We defined an evaluation score, SC, to compare the overlap between clusters based on FEM
291   clustering and real cell types. The number of clusters was related to specific parameters,
292   therefore, each cluster should contain only one cell type as much as possible. Additionally, the
293   cluster's number should be greater than or equal to the number of cell types. Therefore, the
294   clustering parameters were set so that the number of clusters was greater than or equal to the
295   number of actual cell types. For each cluster i, SC_i was measured by the consistency of the cell
296   type within the cluster (the proportion of the cell type with the largest number of cells to the total
297   number of cells in the cluster). We used the following formula to evaluate clustering results:

$$SC_i = \frac{c_{i,max}}{c_{i,T}} \qquad\qquad (3)$$

298       $c_{i,max}$ represents the cell type with the largest number of cells in the i-th cluster. $c_{i,T}$ represents
299   the total number of all cells in the i-th cluster.

300   **Data availability**

301   Publicly available scRNA-seq datasets were used in this study. The PBMCs dataset was
302   downloaded from the 10X Genomics dataset page
303   (https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz).
304   The liver and pancreas scRNA-seq data can be accessed in the NCBI Gene Expression Omnibus
305   (GEO) under accession numbers: GSE81252
306   (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81252), GSE85241
307   (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85241).

308   **Code availability**

309   The FEM software and related scripts have been deposited in the GitHub repository
310   (https://github.com/qingyunpkdd/single_cell_fem).
311
312   **Results**
313   **FEM can separate different cell types**
314   The Reactome pathway gene sets have only 5,741 unique genes and a large amount of gene
315   information was lost when using the FEM algorithm. In comparison GO gene sets had 15,578
316   unique genes. Therefore, GO gene set-based FEM was used for cluster analysis. The Reactome
317   pathway gene sets were used in the analysis of GC-DEF and their cluster results are shown in
318   Figure S1. GO-based FEM results are shown in Figures 3, 4, and 5.
319   Three data sets (PBMCs, liver, pancreas) were used to replace GEM for dimensionality
320   reduction and clustering to verify whether the GO-based FEM algorithm could separate different
321   cell types. The number of clusters was artificially adjusted and the number of clusters was set to
322   be greater than or equal to the number of actual cell types. This was done so that a cluster
323   contained only one main cell type, which allowed us to better distinguish the different cell types.
324   We needed to discover different cell groups and determine the relationship between different
325   cell types for single-cell scRNA-seq data. These cells may have similar functions or a common
326   recent differentiation origin. The FEM method may help determine functionally similar cells.
327   Cells were noted to be different subtypes of the same cell if the two groups of cells were far apart
328   in GEM, but close or partially overlapped in FEM. This result may also indicate that the two
329   groups of cells may perform similar functions.
330   We performed GEM clustering and GO-based FEM clustering on the scRNA-seq data, and
331   compared the two clusters. The FEM clustering results were also compared with real cell types to
332   determine the similarity of cell functions. A low SC score in a particular cluster i indicated that
333   the cluster may contain multiple cell types. A moderate number of a certain cell type A in the
334   cluster i was defined as the main cell type MC_i. For a cluster, all MCs may have the same or
335   similar functions.
336   The results show that the FEM method can distinguish different types of cells that have
337   different functions on three data sets (**Fig. 3, 4, 5**). Figures 3, 4, and 5 show that FEM can find
338   cells with similar origins or different subgroups of a cell. Figure 3 shows three clusters (G1, G2,
339   G3) that have a low SC score and all of the T cells overlapped in cluster G3. The MC_i of NK
340   cells and T cells in cluster G2 literature show that they have a common differentiation origin
341   T/NK cell progenitor. G1's MC_i contained only monocytes (Lim et al., 2013). The FEM in
342   Figure 4 helped determine the relationship of definitive endoderm cells and hepatic endoderm
343   cells (Camp et al., 2017). Figure 5 illustrates that alpha, beta, gamma, epsilon, and delta cells
344   form one cluster in the FEM clusters, and that the cell differentiation tree verifies that they have
345   a common endocrine progenitor (Jacobson & Tzanakakis, 2017).
346   **FC-DEF can directly detect the functional differences between clusters**

347   GC-DEF can reveal the expression of specific functions from the UMAP(t-SNE) diagram of GEM and
348   the results of GC-DEF may aid GEM in determining the functions of a specific group. Official
349   data pre-processing in the PBMC dataset included the removal of cells with excessive
350   mitochondrial genes (> 5%, quality control) and those with too many (> 2,500) or too few (<
351   200) features (Bittersohl & Steimer, 2016). Some cells may have been filtered out in the data
352   pre-processing stage whereas our method did not filter out any cells. Indeed, the filtered cells
353   were found to be located above the NC cells. Using the GC-DEF method, this cell group was
354   found to have a highly expressed cell proliferation-related pathway. Thus, these cells were
355   identified as proliferative (Fig. 6).

356   The most significantly expressed pathway, "hemostasis", was located in the platelet clusters.
357   The "innate immune system" pathway was significantly expressed in the monocyte, DC cell, NK
358   cell, and platelet cell populations, which was consistent with literature results (NORMAN,
359   1995). The top five highly expressed functions were consistent with the cell type characteristics.
360   The other FC-DEF and GC-DEF results are provided in Table S1.

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

365   GO-based GC-DEF results was consistent with the pathway-based methods and literature
366   results (NORMAN, 1995). All other results are presented in Table 3 and Table S1, S2, S3 and
367   S4.

368   **Difference between GO enrichment analysis results based on DEG and FEM**

369   The general GSE analysis was based on DEGs between groups (cell types or clusters). The
370   FEM-based gene set (function) enrichment can be defined as the DEF between groups. The
371   results (Table S5, p-value threshold ≤ 0.05) show that: 1. the enrichment analysis results
372   obtained by the two methods are quite different. 2. There are more GO items based on the FEM
373   method than the method based on DEGs.

374   The possible explanation for the first point is that the two methods are based on different data,
375   and the FEM method is an enrichment result of a single cell. The DEG method is based on
376   results from cell clusters. FEM uses all expressed genes in a cell to calculate GO enrichment,
377   while the DEG method only uses those differentially expressed genes, which may explain why
378   there are more GO items based on the FEM method than the DEGs method. The FEM method
379   focuses on the "differentially expressed GO items", and the DEG method focuses on "the
380   enriched GO items of differentially expressed genes".

381   **Validation with an immune dataset**

382   The Immunologic Signatures Collection (ImmuneSigDB) (Liberzon et al., 2011) was used as a
383   validation dataset to test whether the proposed method could detect sets of genes that had been
384   identified as up-regulated or down-regulated by traditional methods. The ImmuneSigDB is
385   composed of gene sets that represent cell types, states, and perturbations within the immune
386   system (Godec et al., 2016). Figure 7 shows that within the bulk RNA-seq dataset, the cell types

387 from the up-regulated expression marker gene set were highly expressed using the method
388 proposed in this study. This demonstrated the efficacy of the proposed method for detecting cell-
389 type-specific gene sets.
390

Discussion

392 We showed that FEM can be used for cell clustering. It also can merge the GEM method for
393 downstream differential expression analysis to find cell type-specific functions.
394   It was necessary to evaluate the impact of some biological effects on the cell clusters, such as
395 the impact of cell cycles on cell-type clustering results. Within the same cell type, cells in the
396 cell proliferation cycle and non-proliferating cells may be significantly separated in the cluster
397 and UMAP plots, which affects the final cell type identification and differential expression
398 analysis. In some studies, these cells are removed, while they may still be of interest and there is
399 currently no uniform standard for handling these cells. The method proposed in this study
400 directly converts the expression of genes in cells to the expressions of functions and cells can be
401 screened according to their FEM score at any stage of processing.
402   It should be noted that FEM only considers the presence and absence of gene expression,
403 without considering the influences of gene-expression values. Therefore, the proposed method
404 and gene-expression–based methods are complementary and should not be considered as
405 alternatives to one another.
406   We have proven that the FEM-based method can also be applied to cell clustering, and
407 verified the accuracy of cell clustering when compared with real cell types. A direct analysis of
408 the differential expression function between the groups and visually elucidating the differences
409 in specific functions helps to intuitively understand the functions of cells. FEM and GEM were
410 used as two different data sources, while multi-omics data fusion analysis was reused in Seurat
411 and other tools to show the functional results of GEM analysis.
412   The information that each cell can provide may be valuable for single-cell transcriptome data.
413 The same cell type may also be in different physiological states (for example: different stages of
414 the cell cycle, different metabolic states). Cells in a differentiation stage may have a series of cell
415 states. Therefore, it is necessary to develop data analysis tools for individual cells to find those
416 few but possibly important cells, such as those cells that are in a transient state of differentiation.
417

Conclusions

419 The final step in the analysis of gene expression data is to interpret the biological significance of
420 the genes based on a hypergeometric test between the listed genes of interest and the set of genes
421 representing specific functions. If each of the first n genes obtained by differential expression
422 analysis represents a function that does not overlap with the other, then the enrichment analysis
423 will fail. If most genes of a function are expressed in a cell and are screened out in the process of
424 gene selection, this function will also be missed. GSEA at the single-cell level effectively avoids
425 these problems based on the characteristics of single-cell data. The results of the present study

426 showed that direct enrichment analysis at the single-cell level is feasible and yields powerful
427 results.
428   We preliminarily discussed the applicability of the FEM method and the FEM have a wider
429 range of practical applications. The possible applications of the FEM method include:
430 determining the functional differences between different groups; analysing the function of a few
431 outlier cells; analysing the state of a small number of cells on the cell trajectory in the pseudo-
432 trajectory analysis; analysing which cell groups are more similar in function by comparing GEM
433 and FEM clustering; determining marker quality for different cell groups compared with genes
434 (though there is a need for future validation); and more differential functions were found
435 compared with the GEM method. The disadvantages of the FEM method are: the influence of
436 gene expression is not considered; it is impossible to distinguish cell groups with similar
437 functions; and the function of differential expressions and the result of DEG-based enrichment
438 are quite different, and they cannot be substituted for each other.
439

Acknowledgements

441 We would like to thank the Southeast University State Key Laboratory of Bioelectronics for
442 providing the computing resources used in our research.

443

References

445 Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G,
446   Rambow F, Marine JC, Geurts P, Aerts J, Van Den Oord J, Atak ZK, Wouters J, Aerts S.
447   2017. SCENIC: Single-cell regulatory network inference and clustering. Nature Methods
448   14. DOI: 10.1038/nmeth.4463.
449 Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan
450   E, Scholl C, Fröhling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA,
451   Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy
452   S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root
453   DE, Gilliland DG, Jacks T, Hahn WC. 2009. Systematic RNA interference reveals that
454   oncogenic KRAS-driven cancers require TBK1. Nature 462. DOI: 10.1038/nature08460.
455 Bittersohl H, Steimer W. 2016. Intracellular Concentrations of Immunosuppressants. In:
456   *Personalized Immunosuppression in Transplantation: Role of Biomarker Monitoring and*
457   *Therapeutic Drug Monitoring*. Elsevier Inc., 199–226. DOI: 10.1016/B978-0-12-800885-
458   0.00009-6.
459 Blasi T, Buettner F, Strasser MK, Marr C, Theis FJ. 2017. CgCorrect: A method to correct for
460   confounding cell-cell variation due to cell growth in single-cell transcriptomics. *Physical*
461   *Biology*. DOI: 10.1088/1478-3975/aa609a.
462 Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell
463   transcriptomic data across different conditions, technologies, and species. *Nature*
464   *Biotechnology*. DOI: 10.1038/nbt.4096.
465 Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J,
466   Damm G, Seehofer D, Belicova L, Bickle M, Barsacchi R, Okuda R, Yoshizawa E, Kimura
467   M, Ayabe H, Taniguchi H, Takebe T, Treutlein B. 2017. Multilineage communication

468   regulates human liver bud development from pluripotency. *Nature* 546. DOI:
469   10.1038/nature22796.
470   Chen G, Ning B, Shi T. 2019. Single-cell RNA-seq technologies and related computational data
471   analysis. *Frontiers in Genetics* 10:1–13. DOI: 10.3389/fgene.2019.00317.
472   DeTomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. 2019. Functional
473   interpretation of single cell similarity maps. Nature Communications 10. DOI:
474   10.1038/s41467-019-12235-0.
475   Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising
476   using a deep count autoencoder. *Nature Communications* 10:1–14. DOI: 10.1038/s41467-
477   018-07931-2.
478   Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K, Chun J,
479   Kharchenko P V. 2016. Characterizing transcriptional heterogeneity through pathway and
480   gene set overdispersion analysis. Nature Methods 13. DOI: 10.1038/nmeth.3734.
481   Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. 2018. Single sample scoring of
482   molecular phenotypes. *BMC Bioinformatics* 19:1–10. DOI: 10.1186/s12859-018-2435-4.
483   Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: Current state of the science.
484   *Nature Reviews Genetics* 17:175–188. DOI: 10.1038/nrg.2015.16.
485   Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, Mesirov JP, Haining WN.
486   2016. Compendium of Immune Signatures Identifies Conserved and Species-Specific
487   Biology in Response to Inflammation. *Immunity* 44:194–206. DOI:
488   10.1016/j.immuni.2015.12.006.
489   Griffiths JA, Scialdone A, Marioni JC. 2018. Using single‐cell genomics to understand
490   developmental processes and cell fate decisions. *Molecular Systems Biology*. DOI:
491   10.15252/msb.20178046.
492   Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: Gene set variation analysis for microarray
493   and RNA-Seq data. BMC Bioinformatics 14:1–21. DOI: 10.1186/1471-2105-14-7.
494   Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell
495   analysis of differentiation data. *Bioinformatics*. DOI: 10.1093/bioinformatics/btv325.
496   Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: Gene set variation analysis for microarray
497   and RNA-Seq data. *BMC Bioinformatics* 14:1–21. DOI: 10.1186/1471-2105-14-7.
498   Hicks SC, Townes FW, Teng M, Irizarry RA. 2018a. Missing data and technical variability in
499   single-cell RNA-sequencing experiments. *Biostatistics*. DOI: 10.1093/biostatistics/kxx053.
500   Hicks SC, Townes FW, Teng M, Irizarry RA. 2018b. Missing data and technical variability in
501   single-cell RNA-sequencing experiments. *Biostatistics* 19:562–578. DOI:
502   10.1093/biostatistics/kxx053.
503   Hwang B, Lee JH, Bang D. 2018. Single-cell RNA sequencing technologies and bioinformatics
504   pipelines. *Experimental and Molecular Medicine* 50. DOI: 10.1038/s12276-018-0071-8.
505   Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA.
506   2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*.
507   DOI: 10.1186/s13059-016-0888-1.
508   Jacobson EF, Tzanakakis ES. 2017. Human pluripotent stem cell differentiation to functional
509   pancreatic cells for diabetes therapies: Innovations, challenges and future directions. Journal
510   of Biological Engineering 11. DOI: 10.1186/s13036-017-0066-3.
511   Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-
512   cell RNA-seq data. *Nature Reviews Genetics* 20:273–282. DOI: 10.1038/s41576-018-0088-
513   9.

514   Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. 2015. The Technology
515   and Biology of Single-Cell RNA Sequencing. *Molecular Cell*. DOI:
516   10.1016/j.molcel.2015.04.005.
517   Kulkarni A, Anderson AG, Merullo DP, Konopka G. 2019. Beyond bulk: a review of single cell
518   transcriptomics methodologies and applications. *Current Opinion in Biotechnology* 58:129–
519   136. DOI: 10.1016/j.copbio.2019.03.001.
520   Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K,
521   Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-
522   throughput data. *Nature Reviews Genetics*. DOI: 10.1038/nrg2825.
523   Lee E, Chuang HY, Kim JW, Ideker T, Lee D. 2008. Inferring pathway activity toward precise
524   disease classification. PLoS Computational Biology 4. DOI: 10.1371/journal.pcbi.1000217.
525   Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011.
526   Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739–1740. DOI:
527   10.1093/bioinformatics/btr260.
528   Lim WF, Inoue-Yokoo T, Tan KS, Lai MI, Sugiyama D. 2013. Hematopoietic cell
529   differentiation from embryonic and induced pluripotent stem cells. Stem Cell Research and
530   Therapy 4. DOI: 10.1186/scrt222.
531   Luecken MD, Theis FJ. 2019. Current best practices in single‐cell RNA‐seq analysis: a
532   tutorial. *Molecular Systems Biology* 15:e8746. DOI: 10.15252/msb.20188746.
533   McCarthy DJ, Campbell KR, Lun ATL, Wills QF. 2017. Scater: Pre-processing, quality control,
534   normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*. DOI:
535   10.1093/bioinformatics/btw777.
536   McDavid A, Finak G, Gottardo R. 2016. The contribution of cell cycle to heterogeneity in
537   single-cell RNA-seq data. *Nature Biotechnology*. DOI: 10.1038/nbt.3498.
538   McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: Uniform Manifold Approximation and
539   Projection. *Journal of Open Source Software* 3. DOI: 10.21105/joss.00861.
540   Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, van Gurp L, Engelse MA,
541   Carlotti F, de Koning EJP, van Oudenaarden A. 2016. A Single-Cell Transcriptome Atlas of
542   the Human Pancreas. *Cell Systems* 3. DOI: 10.1016/j.cels.2016.09.002.
543   NORMAN P. 1995. Immunobiology: The immune system in health and disease. *Journal of
544   Allergy and Clinical Immunology*. DOI: 10.1016/s0091-6749(95)70025-0.
545   Petegrosso R, Li Z, Kuang R. 2019. Machine learning and statistical methods for clustering
546   single-cell RNA-sequencing data. *Briefings in Bioinformatics* 00:1–15. DOI:
547   10.1093/bib/bbz063.
548   Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M,
549   Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* 177:1888-
550   1902.e21. DOI: 10.1016/j.cell.2019.05.031.
551   Stuart T, Satija R. 2019. Integrative single-cell analysis. *Nature Reviews Genetics* 20:257–272.
552   DOI: 10.1038/s41576-019-0093-7.
553   Tomfohr J, Lu J, Kepler TB. 2005. Pathway level analysis of gene expression using singular
554   value decomposition. BMC Bioinformatics 6. DOI: 10.1186/1471-2105-6-225.
555   Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park JE,
556   Stephenson E, Polański K, Goncalves A, Gardner L, Holmqvist S, Henriksson J, Zou A,
557   Sharkey AM, Millar B, Innes B, Wood L, Wilbrey-Clark A, Payne RP, Ivarsson MA, Lisgo
558   S, Filby A, Rowitch DH, Bulmer JN, Wright GJ, Stubbington MJT, Haniffa M, Moffett A,

559   Teichmann SA. 2018. Single-cell reconstruction of the early maternal–fetal interface in
560   humans. *Nature*. DOI: 10.1038/s41586-018-0698-6.
561   Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: Large-scale single-cell gene expression data
562   analysis. *Genome Biology*. DOI: 10.1186/s13059-017-1382-0.
563   Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, Zhou M, Su J. 2020. Benchmarking algorithms for
564   pathway activity transformation of single-cell RNA-seq data. Computational and Structural
565   Biotechnology Journal 18. DOI: 10.1016/j.csbj.2020.10.007.
566
567
568

Functional gene set

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.

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

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	1.83E-13
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

3

Results of the top five GO-based FEM methods for each cluster (PBMC data set)

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

4

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 (including C and D, the analysis is performed in Seurat). For downstream analysis, FEM can be regarded as another type of GEM. Therefore, Seurat's analysis pipeline can be used to analyse FEM (Figure 1C) or treat FEM and GEM as two different omics data for integrated analysis (Figure 1D). (C) GEM/FEM was used for dimension reduction, clustering, and differential expression genes/function analysis (GC-DEG/FC-DEF). (D) Due to the loss of gene expression information in FEM, the dimensionality reduction and clustering is based on GEM, and differential expression function analysis among clusters in the GEM cluster (GC-DEF) was performed.

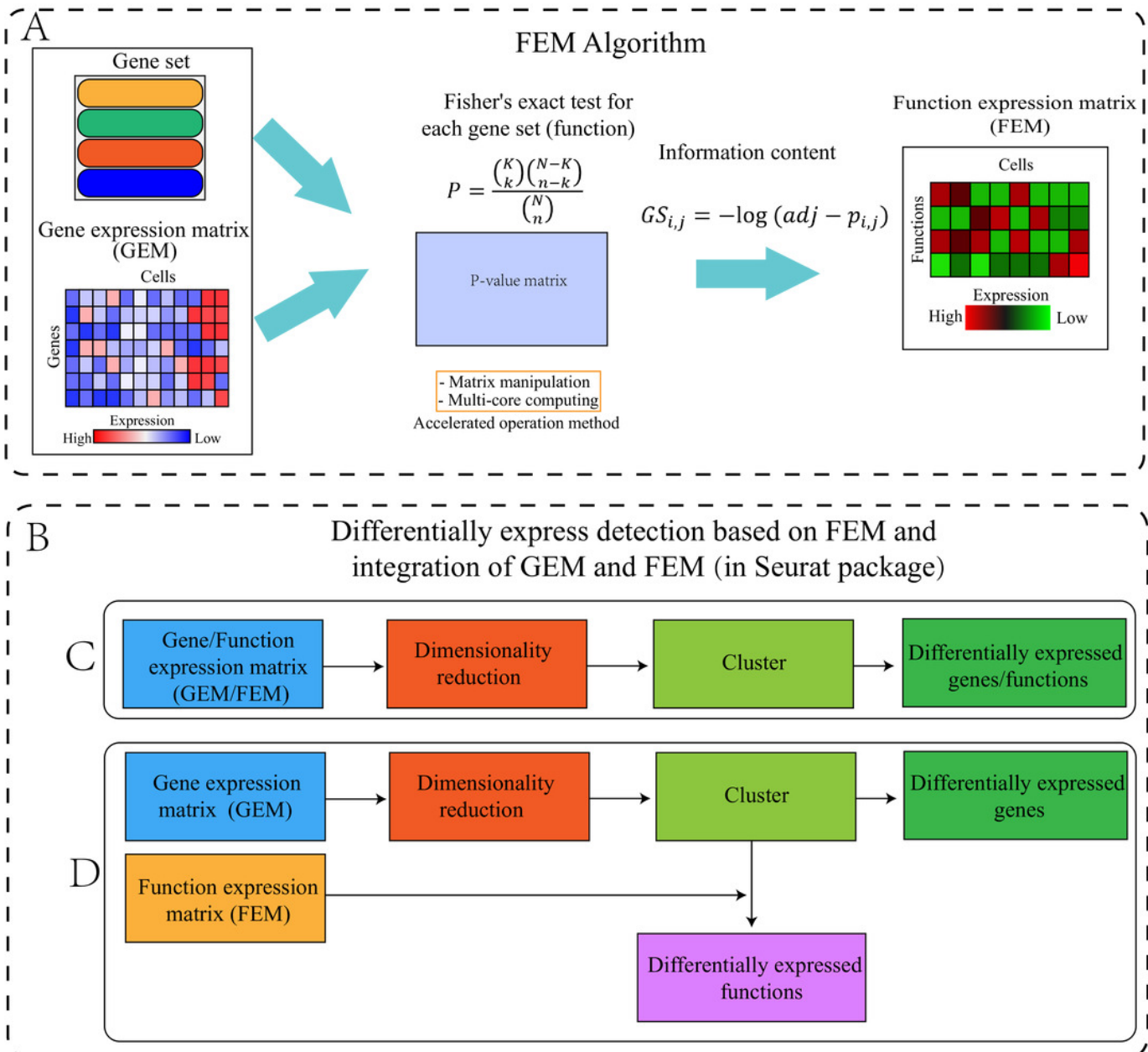
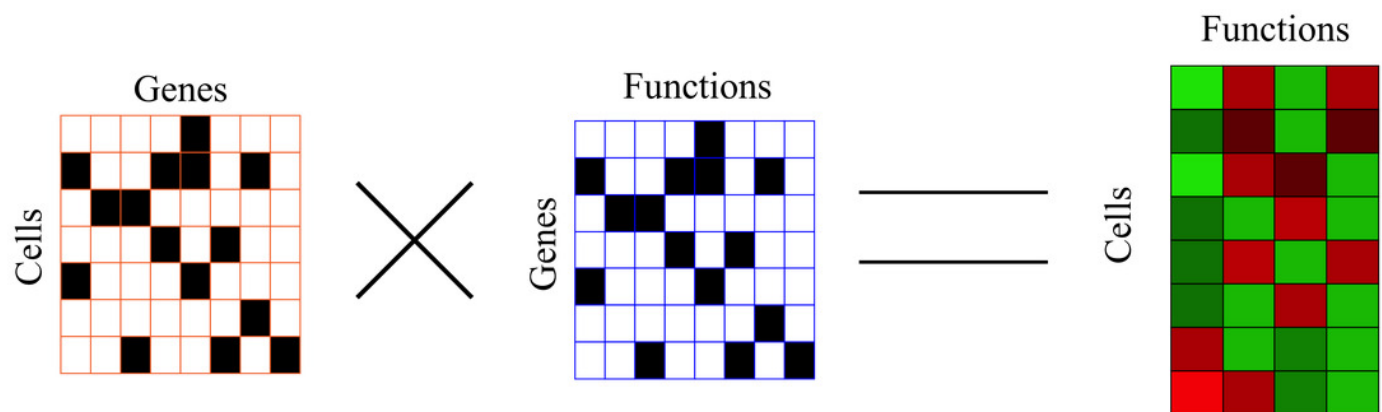



Figure 2

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 <!--[if !msEquation]--> <!--[if !vml]--> <!--[endif]--><!--[endif]--> and set <!--[if !msEquation]--> <!--[if !vml]--> <!--[endif]--><!--[endif]-->. (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 a multi-process operation.

A

Matrix manipulation

B

Multi-core computing

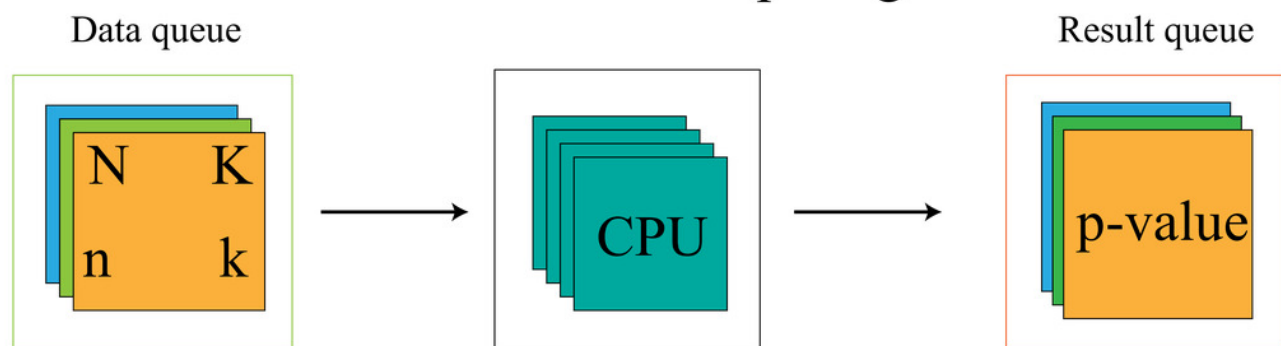



Figure 3

The comparison between GEM clustering (3A) and GO-based FEM clustering (3B) helps to find the similarities between immune cells.

(A) The results of GEM clustering. (B) The GO-based clustering results. (C) The correspondence between each cluster and cell type of GO-based clustering. (D) The differentiation tree of immune cells (Lim et al., 2013). All clustering results are plotted by UMAP. Figure C shows that there are three clusters with low SC scores. The G1 cluster is mainly composed of two types of monocytes (CD14+_Mono, FCGR3A+_Mono). It can be seen from Figure 3D that all monocytes have the same recent developmental origin. G2 is composed of NK cells and CD8 T cells. Figure 3D shows that both NK cells and T cells are developed from T/NK cell progenitor cells. G3 is composed of naive CD4 T, CD8 T, and memory CD4 T. The differentiation tree in Figure 3D shows that all T cells have a common nearest ancestor. The No. 7 cluster has a low score, however, it has a small number of cells and the number of cells in all cell types is not the main cluster of the corresponding cell type, and is ignored.

A

GEM-based cluster

B

FEM-based cluster

C

cell_type\cluster	4	3	0	2	7	5	1	8	6
B	320	0	2	0	8	0	8	0	12
Naive_CD4_T	2	0	418	0	3	0	134	0	127
CD14+_Mono	0	317	0	175	0	0	0	0	0
CD8_T	1	0	77	0	3	115	114	0	74
NK	0	0	0	0	12	135	4	0	3
DC	0	1	0	29	5	0	0	0	0
FCGR3A+_Mono	0	18	0	143	0	0	0	0	0
Platelet	0	0	0	1	0	0	1	11	0
Memory_CD4_T	1	0	116	0	12	3	289	0	6
SC	0.99	0.94	0.68	0.5	0.28	0.53	0.53	1	0.57

G1   G2   G3

D

T /NK cell progenitor

granulocyte-monocyte progenitor (GMP)

T cells   Natural killer cells

Monocyte

Figure 4

Comparison of GEM clustering (4A) and GO-based FEM clustering (4B) helps to find the similarities between liver cells.

(A) The results of GEM clustering. (B) The GO-based clustering results. (C) The correspondence between each cluster and cell type of GO-based clustering. (D) The developmental trajectory of stem cells (Camp et al., 2017). All clustering results are plotted by UMAP. Figure 4C shows that a cell group has a low SC score. G1 is mainly composed of definitive endoderm and hepatic endoderm cells. These two types of cells are far apart in GEM (Figure 4A). But it is close to each other in GO-based FEM cluster. From Figure 4D, we find that these two cells are adjacent to each other on the cell differentiation trajectory, and they are the progenitor cells of other cells in the liver.

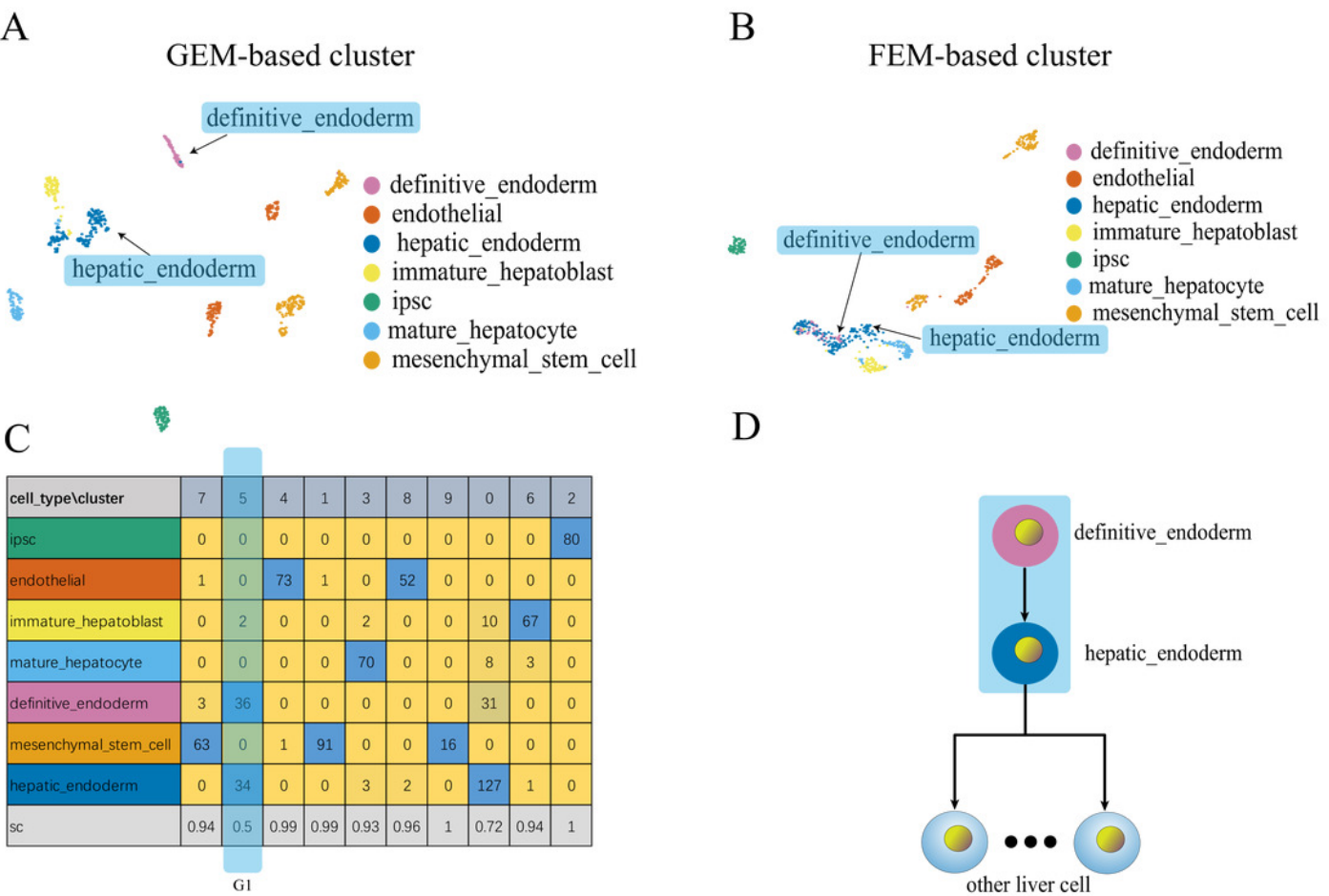



Figure 5

Comparison of GEM clustering (5A) and GO-based FEM clustering (5B) helps to find the similarities between pancreatic cells.

(A) The results of GEM clustering. (B) The GO-based clustering results. (C) The correspondence between each cluster and cell type of GO-based clustering and the SC score. (D) The differentiation lineage of pancreatic cells (Jacobson & Tzanakakis, 2017) . All clustering results are plotted by UMAP. Figure 5C shows that the two cell groups have the lowest scores. Although G1 has a low score, it has a small number of cells. The number of cells in all cell types is not the main cluster of the corresponding cell type, so it will be ignored. G2 is composed of five types of cells (alpha, beta, delta, gamma, epsilon). Although the epsilon cell has only two cells in cluster 1, but it has only three cells, so it is considered that it is distributed in G2. These five types of cells are far apart in the GEM UMAP plot, but close to each other in the GO-based FEM plot. Figure 5D shows that all these cells have a common progenitor cell differentiation.

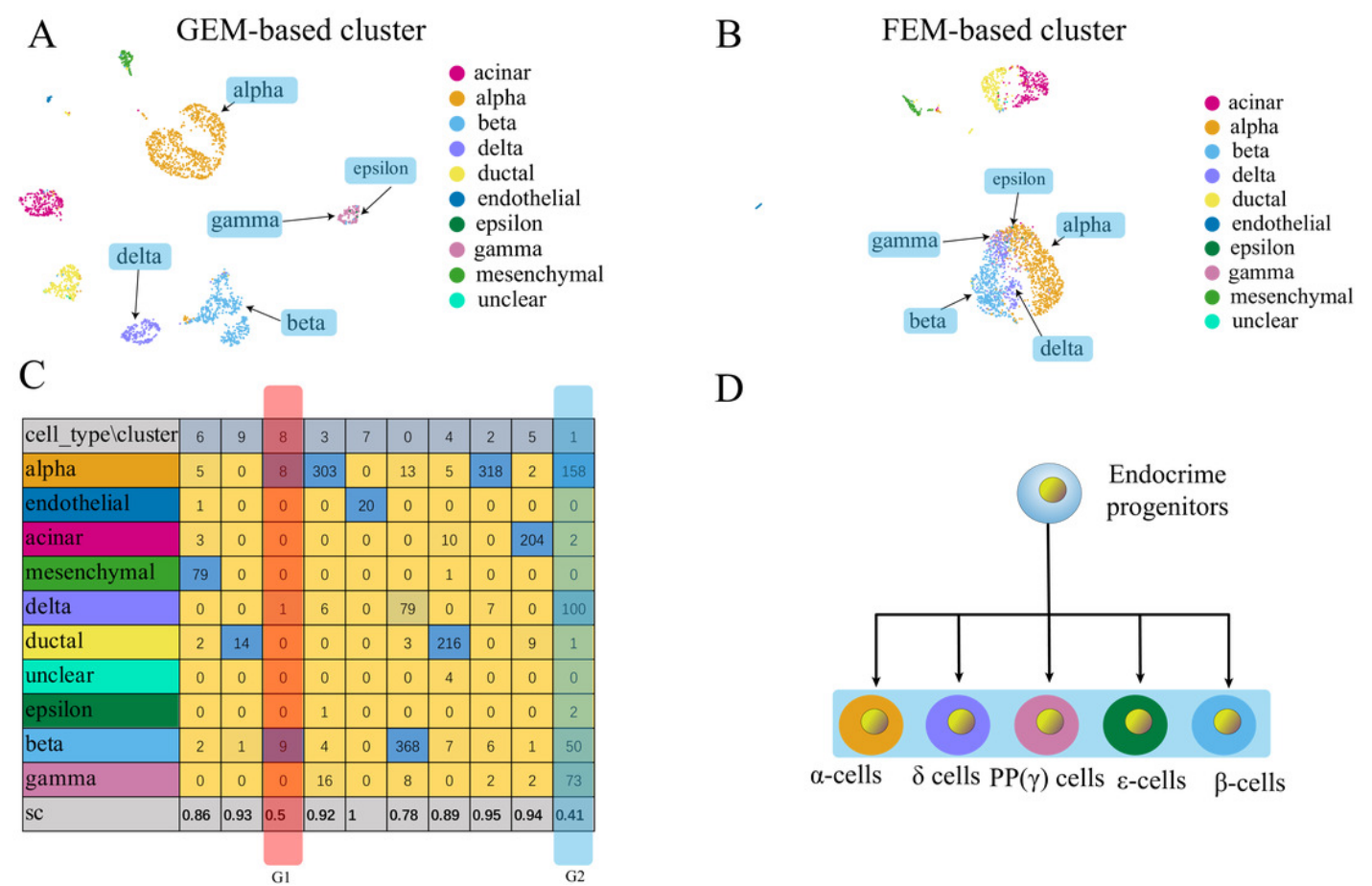



Figure 6

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 (when compared with Figure 5A). (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. For downstream analysis, cell screening may result in some cells being filtered out, which may hold some interest. In the process of biological research, cells at special stages often account for only a small part of the total number of cells, but they may be important for explaining biological processes. For example, a small part of cells in the cell cycle stage (Figure B). They are clustered in the NK cell group in clustering, but compared with other NK cells, they are a prominent part of the UMAP plot. When doing differential expression between cell groups, the signals of these proliferating cells will be masked by NK cells. The FEM method is based on a single cell, so it can directly display the functions of these special cells even at single cell level.

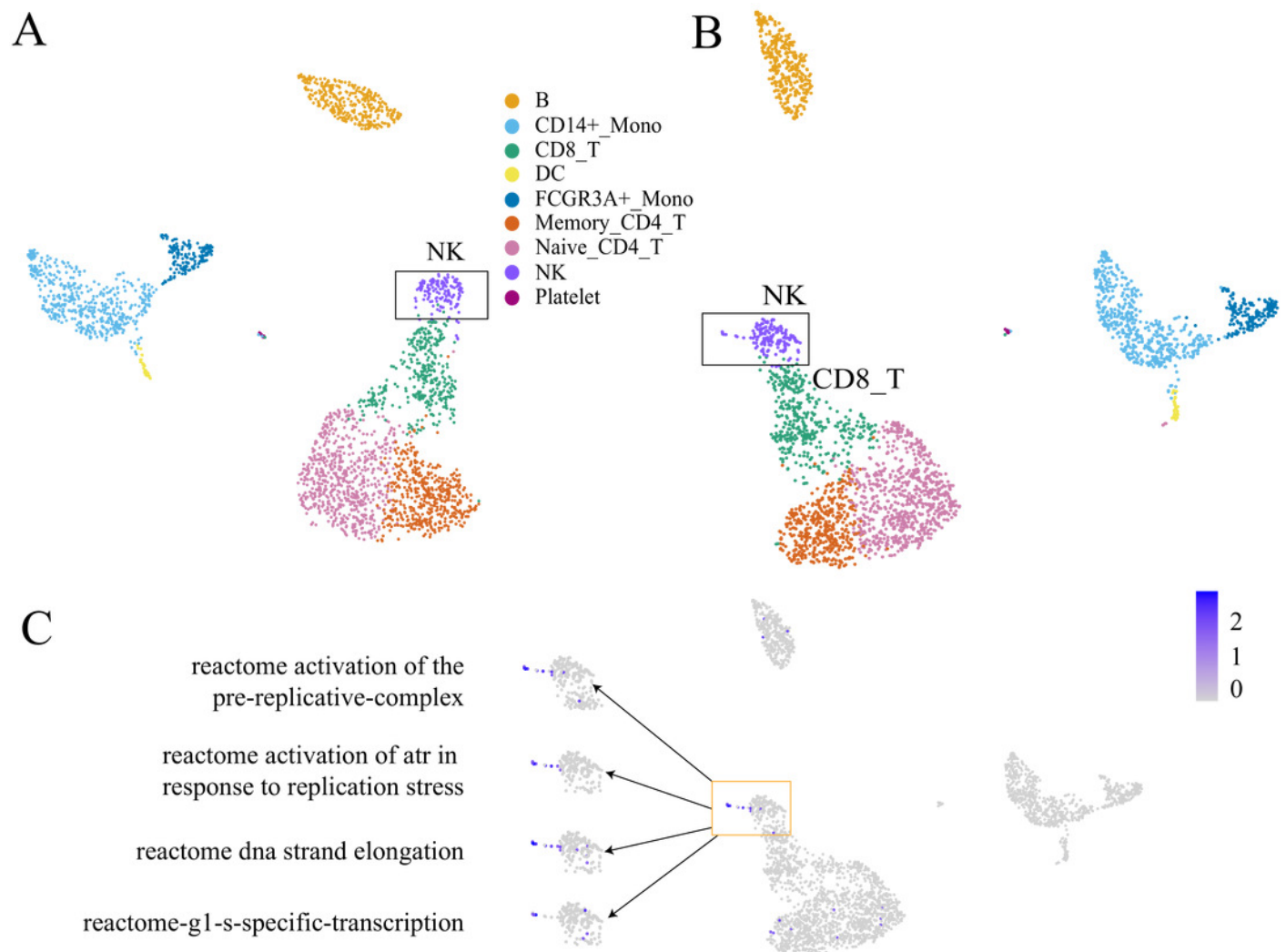



Figure 7

Validation results of Immunologic Signatures Collection gene sets.

ImmuneSigDB (Godec et al., 2016) is a collection of gene sets of human and mouse immune cells collected from the GEO database. All data are manually corrected. Each gene set represents genes that are highly expressed in a specific immune cell type compared to another or all other cell types. Here, we verify whether the expression of each gene set (in this case, functional expression) is consistent with the corresponding cell type. (A) UMAP plot for each cell type based on GEM. (B-G) The expression values of various cell type-specific gene sets in each cell. The figures show that the functional expression obtained by our method is consistent with the corresponding cell type.