Enhancing single-cell cellular state inference by incorporating molecular network features

Ji Dong[1,3,*], Peijie Zhou[2,*], Yichong Wu[2,*], Wendong Wang[1,*], Yidong Chen[1,,3,*], Xin Zhou[1,*], Haoling Xie[1,3], Jiansen Lu[1,3], Xiannian Zhang[1,3], Lu Wen[1,3,], Wei Fu[1,#], Tiejun Li[2,#], Fuchou Tang[1,3,4,#]

[1] Beijing Advanced Innovation Center for Genomics (ICG), College of Life Sciences, Department of Obstetrics and Gynecology, Third Hospital, Peking University, Beijing 100871, China

[2] LMAM and School of Mathematical Sciences, Peking University, 100871 Beijing, China

[3] Biomedical Institute for Pioneering Investigation via Convergence and Center for Reproductive Medicine, Ministry of Education Key Laboratory of Cell Proliferation and Differentiation, Beijing 100871, China

[4] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

[*] These authors contributed equally

[#] Correspondence: fuwei@bjmu.edu.cn (W.F.), tieli@pku.edu.cn (T.L.), tangfuchou@pku.edu.cn (F.T.),

Abstract

We present SCORE, a network-based method, to simulate the dynamic changes of molecular networks among different cellular states (https://github.com/wycwycpku/RSCORE). Using published and newly generated single-cell sequencing data, SCORE outperforms existing methods in accuracy, robustness, scalability for better data integration and batch effect removal. Moreover, SCORE can identify crucial gene modules and construct the characteristic molecular interaction networks for a cellular state, providing more biological insights rather than mainly statistical interpretation and annotation.

Introduction

Instead of performing function in isolation, genes tend to form complex molecular networks and function in conjunction to affect the cellular or organismal phenotypes[1,2]. Although it has been recognized for a long time that molecular networks are dynamic during organismal development and differentiation, how to simulate this process is still challenging.

With the rapid development of single-cell RNA sequencing (scRNA-seq) techniques, nowadays, researchers can easily profile the transcriptome of large number of cells at single-cell resolution, and identify new cell types and intermediate cellular states within a certain organismal system[3-5]. To fully utilize these rich datasets, many efficient computational methods have been developed, such as, Seurat, SCANPY and SINCERA[6-8]. However, these methods usually calculate the transcriptional similarity inferred from individual genes to perform clustering of cells and detect the marker genes that can uniquely distinguish the identified cell types, while ignoring the integrated and synergistic nature of molecular interaction networks, which in fact formulate the specific cell fates and provide more biological insights. In addition, the performance of individual gene-based methods can be influenced by the prevalent dropout events and other experimental artifacts in single-cell sequencing data, and the performance strongly depends on the subjective selection of highly variable genes in the pre-processing step.

Recently, several methods such as SCENIC, PAGODA and CSN, proposed to identify cell types by incorporating some biology knowledge-based information, which utilized transcription factor-based regulatory networks, functional modules and cell type-specific networks to facilitate the biological interpretation, respectively[9-11]. However, due to the strong technical noises there are still no optimal methods to accurately infer the gene-gene or cell-cell relationship from the sparse scRNA-seq datasets[12]. In the meantime, accumulating public data on molecular interactions derived from sound experimental or computational evidences can provide rich prior

biological knowledge to reduce the false-positive rate of molecular network inference [13,14]. For example, with the aid of protein-protein interaction (PPI) databases, SCENT is able to infer the single-cell entropy, while netSmooth can temper the noisy scRNA-seq data[15,16].

We hypothesize that during the organismal development and differentiation, the transition from a cellular state to another is accompanied by the destruction of critical molecular networks and the reconstruction of novel ones (Figure 1A). In view of this, we introduce a new method, **SCORE** (**S**ingle-**C**ell m**O**lecula**R** n**E**twork), to simulate the dynamic changes of molecular networks from scRNA-seq datasets by incorporating the experimentally validated and high-confidence molecular interaction information from public databases. We validated the accuracy, robustness and scalability of SCORE to uncover cell states using gold-standard scRNA-seq datasets. In addition, with **SCORE**, we accurately integrated five human embryonic datasets and analyzed a newly generated human adult small intestine dataset sampled from six patients.

Results

1. The workflow of SCORE

In brief, as we reason that genes perform cellular functions through interacting with other genes in molecular networks, we need to extract the molecular networks that undergo dynamic changes among different cellular states (Figure 1B). Thus, we first generate a weighted gene-gene Pearson correlation network. For scRNA-seq datasets, Pearson correlation coefficient can cover more authentic gene-gene relationships compared with other algorithms, although it will also introduce large amount of false positives. To reduce the false positive rate, we use a curated PPI network to trim the gene-gene relationships; likewise, data-irrelevant interactions of PPI network are pruned by the weighted correlation network. The obtained weighted network is then decomposed into numbers of small molecular networks, termed as modules, by random walk algorithm. Finally, each module is scored within each cell using an AUC based algorithm[9], and a cell-module matrix is obtained, which represents the activity of individual module within each cell.

Next, this cell-module matrix can be utilized to perform downstream analyses such as visualization, clustering, and cell lineage analysis. Importantly, inspired by the concept of Steiner Tree in graph theory, SCORE also constructs the characteristic molecular interaction network (CMIN) to annotate a certain cellular state (see Methods for more details). For the convenience of users, SCORE is seamlessly compatible with the popular R package Seurat[6].

2. Evaluation of the accuracy, robustness and efficiency of SCORE

To verify the rationale and assess the performance of the algorithm, we applied **SCORE** to a gold-standard scRNA-seq dataset for cell clustering analysis, with 561 cells derived from seven human cell lines (GSE81861) after quality control. Two experimental batches exist in the cell lines for GM12878 lymphoblastoid cells and H1 embryonic stem cells. In the previous literature[17], several clustering methods were benchmarked without providing the additional information on batch identity, and the

reference component analysis (RCA) achieved the highest adjusted rand index (ARI) of 0.91, while other methods showed inferior performance (All=HC: 0.66; HiLoadG-HC: 0.53; BackSPIN: 0.64; RaceID2: 0.15; Seurat: 0.70). Here, we intend to continue the benchmarking in the same set-up, by comparing the performance of SCORE with other state-of-art feature extraction methods based on genes regulatory network, and highlight the strength of SCORE in the accurate identification of the cell lines and the effective alleviation of undesired experimental noises.

As shown in Figure 2A, SCORE yielded highest accuracy, with the ARI amounting to 1.00 (clustering by SNN[18]) and 0.99 (clustering by SIMLR[19]) respectively. Both the t-Distributed Stochastic Neighbor Embedding (t-SNE) plot and SC3[20] similarity matrix demonstrated that the batch discrepancy was removed within the same cell line, while the sharp distinction among different cell lines was still retained. In comparison, the direct clustering based on expression matrix confronted with limitations, by blurring the distinct cell identities (e.g. K352 and H1437 or IMR90 cells) and discriminating the unwanted experimental batches in both GM12878 and H1 cell lines (indicated by the separate clusters in t-SNE plot and independent blocks in SC3 matrix in Figure 2A). Despite that the cell-specific network (CSN) method separated different cell lines, clustering variations induced by technical artifacts seemed to be strengthened. Clustering guided by SCENIC continuous features generated relatively satisfactory results comparable to (but still inferior than) the analysis by SCORE, however with considerable computation costs.

In addition, the performance of SCORE was robust under the selection of input variable genes for analysis. As shown in Figure 2B and S1, SCORE always outperformed other methods for all the selected gene numbers ranging from 2,000 to 14,000, both in terms of clustering ARI and the removal of experimental batch effects, as indicated by the SC3 similarity matrix and t-SNE plot. As long as the gene number exceeded 2,000, SCORE could robustly recover the cell line identities with optimal resolution. In comparison, while SCENIC achieved satisfactory clustering results with

5,000-11,000 genes, the clustering ARI was significantly reduced for the case of 14,000 genes due to the severe batch effects.

The implementation of SCORE is also efficient, mainly due to the fast module decomposition of PPI by random walk distance, and the highly parallelizable calculation of module activity matrix. With 14,000 top variable genes, SCORE finalized the analysis within 2 minutes, while the implementation of SCENIC lasted for 6 hours in R. In fact, from the running time tests over various groups of genes (Figure 2B), the implementation of SCORE was typically 10-20 times faster than CSN and 100-400 times faster than SCENIC. Notably, SCORE could handle the task of 15,000 cells and 14,000 genes in less than 10 minutes (Figure 2C).

Overcall, application of SCORE to the gold standard dataset validates our rationale, that incorporation of molecular interaction network in single-cell transcriptome analysis can facilitate the accurate cell identity dissection and reduce the technical artifacts in experiments. The comparisons with other methods also suggest the capability and potential of SCORE for the unbiased and efficient analysis and integration of large-scale transcriptome datasets, which will be explored below.

3. Integration and multi-scale comparison of five human embryonic datasets

Using single-cell RNA-seq technique, many studies have uncovered unprecedented findings in terms of human embryonic development. However, most of these studies focused on just one organ or tissue and failed to consider the human embryonic development in a holistic view. In addition, it is also a challenge for existing methods to integrate the datasets sampled from different organs and different time points because of batch effects introduced by varied dissociation protocols, organ-specific differences, etc. Given this, we used SCORE to integrate five high-quality human embryonic datasets previously generated by our group, including fetal germ cells (FGCs)[21], heart[22], kidney[23], prefrontal cortex[24], and cerebral cortex[25], spanning from 4 to 26 weeks of fetal development. Importantly, we re-organized the five datasets using

uniform pipeline and format, which provided a rich and convenient resource for studying human embryonic development (https://github.com/zorrodong/HECA). To evaluate the integration result, we utilized the mixability of common cell types shared by these datasets, such as immune cells, erythroid cells and endothelial cells across different organs. Notably, SCORE obtained a reasonable result with cell types grouped by their own identities, while Seurat pipeline on individual gene expression matrix failed to cluster these common cell types (Figure 3A-C). We also used two batch effect correction methods, Harmony[26] and CCA[27], to integrate these datasets; however, they all exhibited defective results. Even worse, both Harmony and CCA introduced serious artificial results and unrelated cells were wrongly grouped together.

In addition, we also analyzed each of these five datasets separately (Figure S2-S6). Take the human embryonic kidney dataset as an example, standard clustering using Seurat pipeline showed a chaos state, especially in renal interstitium (RI) (Figure S2). Specifically, there was a distinctly isolated cell population of 19-week fetal renal cells, which might be caused by batch effects. In contrast, SCORE presented a much better result. All fetal cells were first separated into renal cell populations and non-renal cell populations, including immune cells (IMMs), erythrocytes (ERs), and endothelial cells (EDs). Renal cells comprising glomerular cells, renal capsule cells and renal tubular cells were then classified into 9 clusters according to the anatomic structure of nephron, namely, cap mesenchyme (CM), podocytes (PDs), proximal tubule (PT), loop of Henle (LH), distal convoluted tubule (DT), collecting duct (CD), intraglomerular mesangium (MG), extraglomerular mesangium (EM), renal interstitium (RI). Importantly, the umap plot using SCORE displayed an accurate developmental trajectory of nephrons from the cap mesenchyme to the formation of epithelial tubules (Figure S2). For other four datasets, SCORE also performed better than Seurat. In heart dataset, SCORE firstly divided cardiomyocyte cells (CMCs) into compact CMCs and trabecular CMs, and then portioned compact CMCs into atrial CMCs (CMCs-A) and ventricular CMCs (CMCs-V), which was consistent with the

anatomic structures of heart (Figure S3). In two cortex datasets and germ cell dataset, slight batch effects were detected in Seurat results but not in SCORE results (Figure S4-6).

4. Expression landscape of human adult small intestine

Compared with embryonic datasets, adult datasets tend to be more complex. Besides, there are more individual differences in adult datasets, which is a major influence factor causing confusion and inaccuracy when performing scRNA-seq analysis. To demonstrate that SCORE can eliminate these unwanted effects, we generated a new human adult small intestine dataset from the normal adjacent tissues of six small intestine cancer patients (Figure 4A). Using SCORE, we identified nine clusters, which corresponded to nine cell types, namely, *OLFM4*+ epithelial progenitor cells, enterocytes, goblet cells, B cells, T cells, macrophages, mast cells, fibroblasts and endothelial cells. The differentially expressed genes (DEGs), differentially activated modules (DAMs) and the cell distribution in umap plot all supported the clustering accuracy (Figure 4B). However, Seurat pipeline grouped cells according to the patient information, which indicated the existence of batch effect and would result in serious consequence for downstream analyses. For example, when using Seurat pipeline, enterocytes and B cells of different patients clearly formed independent clusters (Figure S7).

In addition, we used SCORE to construct the steiner tree network for each cluster combining all the differentially expressed modules and genes of the given cluster (Figure 4C; Figure S8). Interestingly, the network of *OLFM4*+ epithelial progenitor cells showed a different pattern compared with those of two differentiated cell types, enterocytes and goblet cells. The network of *OLFM4*+ epithelial progenitor cells did not have evident branches, indicating a high level of signaling promiscuity and a high uncertainty, while the networks of two differentiated cells showed clear structuralization, implying a low uncertainty and the commitment of certain cell lineages. This result is consistent with the signaling entropy concept that multipotent

stem/progenitor cells possess multi-lineage differentiation potentials and should not prefer any particular cell lineage, thus tend to be at an uncertain state[15]. In contrast, differentiated cells are committed to particular lineages, thus tend to reduce the uncertainty.

Discussion

The close cooperation of different genes forms gene modules to fulfill specific cellular functions, and a certain cellular state or cell type can be well depicted by the activities of various gene modules[1]. As cellular states transit rapidly during organismal development and differentiation, how to simulate these dynamic processes and uncover the corresponding cell fates becomes a biologically fundamental issue. In this study, we present a new computational method, SCORE, to infer this dynamic change and reveal cell development trajectory from the point of view of molecular network. There are several advantages of SCORE compared with currently widely used methods.

Firstly, the hypothesis of SCORE is more biologically reasonable. Genes function in conjunction through molecular networks rather than in isolation. However, most published methods ignored the importance of molecular networks and just calculate the transcriptional similarity inferred from individual genes. A novel method called SCENIC utilizes transcription factor (TF) based networks to retrieve regulatory activity patterns, which can be used to annotate cellular states. However, SCENIC only measures the activity patterns of approximately 1500 TFs, and a recent evaluation showed that the sensitivity of SCENIC is only about 5% (about 75 TFs)[28]. Thus the resolution of SCENIC is limited, especially when analyzing highly heterogeneous single-cell data. In contrast, SCORE infers the molecular network from most of the expressed genes, and uses the network signatures rather than individual genes to define a certain cellular state, which can improve the resolution and accuracy.

Secondly, SCORE can significantly reduce the false positive rates and yield more accurate results. SCORE uses the curated PPI network to correct the inferred gene-gene relationship, and all the interactions are literature-proved. As shown in Figure 1, compared with other popular methods, SCORE possesses higher accuracy and robustness.

Thirdly, SCORE can overcome dropout effects and other technical variations to successfully integrate different datasets. One of the major drawbacks in scRNA-seq technique is the high dropout rate. Since SCORE performs the analyses based on gene modules rather than individual genes, it can effectively reduce the unwanted dropout effect. In addition, PPI corrected correlation and AUC based module score are both able to eliminate the batch effects. As shown in Figure 2 and 3, SCORE successfully integrated five human embryonic datasets and one human adult small intestine dataset sampled from six patients, while other methods failed.

Fourthly, in addition to the DEGs, SCORE can also identify the DAMGs in differentially activated gene modules and infer their relationships through the constructed CMIN for each cell fate (Figure 3). Moreover, due to the flexible framework of SCORE, it can be easily applied to deal with other networks, such as pathway network, metabolism network, etc.

Finally, the molecular network decomposition step in SCORE workflow is realized with high efficiency and independently of data size. The major computational cost of SCORE lies in the quantification of module activities by AUCell, which scales linearly with cell numbers and has been easily paralleled in the R implementation. Thus, SCORE is of high scalability and is able to cater to the increasing size of scRNA-seq datasets nowadays.

However, as SCORE relies on the curated PPI network, it may be not capable to handle the datasets of organisms without high-confidence molecular interaction information. Besides, when the available gene number is too low, the accuracy of SCORE may be impaired. Thus, we highly recommend researchers apply SCORE to analyze high-quality datasets with greater sequencing depth.

In summary, with the high accuracy, robustness, and scalability, SCORE can help us

to explore the scRNA-seq datasets in a more biologically reasonable manner, and gain more insights into the complex biological systems.

Figure legend

Figure 1. The workflow of SCORE

(A) SCORE assumes that the cellular state transition is associated with the activation/inactivation of functional modules of molecular interaction network, which can be inferred from single-cell transcriptome data.

(B) In the SCORE workflow, the input protein-protein interaction (PPI) network from public database and gene correlation inferred from single-cell dataset are trimmed to construct a weighted molecular interaction network (WMIN). The random walk approach then is applied to WMIN to decompose molecular interaction modules via a consensus strategy, and the activation score of modules for each cell is calculated from AUCell. Downstream analysis are performed based on the obtained cell-module activity matrix to cluster and visualize the cells against technical variations, and construct the characteristic molecular interaction network (CMIN) for each cell state.

Figure 2. The performance assessment of SCORE.

(A) The performance of SCORE on the gold-standard cell-line benchmarking dataset was compared with the direct analysis on raw expression matrix (denoted as Raw) and other gene regulatory network (GRN) based methods (i.e. CSN and SCENIC with continuous/binary features). SCORE outperforms other methods in terms of both clustering accuracy (indicated by Adjusted Rand Index: ARI of different clustering approach SNN and SIMLR, as well as the batch removal effect revealed by the SC3 similarity matrix and t-SNE plot) and run time. The number of input highlight variable genes is set as 5000, and the colors in t-SNE plot denote cell lines identity collected in experiment.

(B) The assessment of SCORE robustness to feature selection in gold-standard dataset and comparison with other methods. SCORE always achieves the highest ARI and shortest run time regardless of the number of selected HVGs.

(C) The scalability of SCORE with the increase of input genes and cells number, tested by down sampling of the integrated human embryonic datasets. The run time of SCORE scales almost linearly with the number of cells, and does not witness the

significant increase as genes number exceeds 11,000. The implementation of SCORE completes within 10 minutes with 15,000 cells and 14,000 genes.

Figure 3. Unbiased integration of five human embryonic datasets.

(A) The umap visualizations of different data integration results implemented by SCORE, direct merge of datasets (Raw), Harmony and CCA (aligned from left to right in each row). The cells are colored by organ information (the top row), embryonic week information (middle) and clustering results based on integrated data (bottom) respectively.

(B) The umap visualizations of data integration results by SCORE and other methods, with cells colored by the expression level of marker genes for immune (top), erythroid (middle) and endothelial (bottom) cells. The cells that highly express the marker genes are denoted by the red circles.

(C) Comparison of marker genes analysis based on different data integration results by SCORE and other methods, with the top 10 marker genes displayed for each cluster. The columns in the matrices represent cells, and the rows represent genes. The wrongly clustered cells, which do not significantly express marker genes of the assigned cluster, are marked by the black squares.

Figure 4. Analysis of human adult small intestine dataset by SCORE.

(A) The umap visualizations based on SCORE (top row) or raw expression matrix (bottom row). The cells are colored by patient information (left column) or SCORE clusters (right column) respectively. The cells subject to batch effect in the raw expression matrix analysis are marked by the red circles. The SCORE clusters are annotated with the cell type labels from the marker genes information.

(B) The marker genes analysis based on the SCORE clustering results, with the top 5 marker genes displayed for each cluster. The names of marker genes are colored according to the SCORE clusters in (A).

(C) The characteristic molecular interaction networks (CMIN) for the epithelial progenitor cells (left) and the differentiated enterocytes (middle) and goblet cells (right) discovered by SCORE. The nodes sizes in CMIN represent the Pagerank score of each gene and the genes of top 20 Pagerank scores are displayed with their names in CMIN. The nodes colors denote the classification of genes as both DAMGs and DEGs (red), DAMGs (green), DEGs (orange), and the connecting Steiner genes (gray) in CMIN.

Methods

Overview of SCORE

For the convenience of users, SCORE is seamlessly compatible with the Seurat pipeline. The input for SCORE workflow includes a single-cell gene expression matrix and a PPI network. Both the cells and genes in the expression matrix could be pre-filtered by users, while it is highly recommended that enough number of genes (>5000) should be retained to achieve more robust analysis of SCORE, as shown via the gold-standard cell line dataset (see main text and SI). The nodes of PPI network should overlap considerably with the gene names in the expression matrix, and the edges of the network shall represent the corresponding molecular interactions with relatively high confidence. In the R implementation of SCORE, the procedure supports the automatic download of PPI from public database such as Biogrid and STRING, which is recommended in the standard workflow. The output of SCORE is a cell-module matrix representing the activity of individual dynamic module within each cell, which can be utilized for downstream visualization, clustering and cell lineage analysis.

Dissection of dynamic molecular networks

To simultaneously reduce the false-positive interactions from correlation inference, and prune data-irrelevant interactions of PPI network, SCORE constructs a weighted molecular interaction network (WMIN) $\mathcal{G}(V, E, W)$ by combining the data-driven and knowledge-based approaches. The vertex set V of the network only consists of the genes in the input expression matrix, and the edge set E obtained from the input PPI network. The weight W_{ij} on the edge E_{ij} is the Pearson correlation coefficients between the corresponding nodes i and j calculated from the single-cell gene expression matrix. To improve the interpretability of molecular networks and highlight the co-expression features, by default we only keep the edges with positive weight in the WMIN.

To achieve fast and robust identification of dynamic molecular networks, SCORE

utilizes the consensus detection of the weighted network community through random walk approach. Given the weighted network $\mathcal{G}(V, E, W)$, a random walk on the network is naturally induced, whose transition probability matrix (TPM) P is defined by

$$P_{ij} = \frac{W_{ij}}{d_i}, \ \ d_i = \sum_{j \in \mathcal{N}(i)} W_{ij}.$$

SCORE constructs an ensemble of random walks with different step lengths on the WMIN, with the TPMs $P^{l_1}, P^{l_2}, \ldots, P^{l_R}$, where the power of matrix l_m denotes the length of time step. The walktrap algorithm is applied to detect network community respectively for each random walk. The algorithm partitions the network based on the distance induced by the random walk, based on the intuition that the random walker will be "trapped" in the closely-connected sub-networks (termed as modules). All modules with molecule number larger than 3 in each run will be kept as the final dynamic modules, resulting in the module set $\{M_k\}_{k=1}^{L}$. It would be possible that certain modules occur repeatedly in different run of walktrap algorithm, indicating their stability to form a closely-connected community. SCORE strengthens the weight of such modules automatically in the subsequent analyses, by restoring all the modules without deletion of repeated items.

Quantification of the module activity

For each detected module, SCORE utilizes AUCell to quantify its activation level within each individual cell. Given cell x, genes are ranked in descending order according to their expression level in x. By default, the z-score is adopted in SCORE to rank the genes in order to remove the effect of scaling. The recovery curve (ROC) for module M_k is then derived by counting the top ranked genes enriched in M_k. The activity measure $A_k(x)$ of module M_k in cell x (consists of the final output matrix of SCORE) is defined as the area under the curve (AUC) for the top ranked genes. Intuitively, modules with higher activity in the biochemical process tend to possess the high-ranking gene expression level, therefore associate with higher activity measure. The AUCell procedure is independent of gene expression unit or

normalization method, therefore achieving the effective removal of batch effect in the single-cell experiments.

Downstream Analysis

The obtained module activity $A_k(x)$ matrix from SCORE (whose rows represent modules and columns represent cells) can replace the raw gene expression matrix as the input for downstream analysis, such as dimension reduction, clustering and lineage inference. As shown in the various datasets of the main text, the downstream analysis based on SCORE module activity features outperforms the raw expression matrix, in terms of clustering accuracy, development lineage trend and removal of experimental batch effects. Hence, the workflow of SCORE can be understood as the extraction of biologically meaningful and robust features, guided by molecular interactions in the single-cell transcriptome data.

For the convenience of downstream analysis, the R implementation of SCORE is deeply fused with the workflow of Seurat v3.0 package. The input expression matrix to SCORE can be a Seurat object with RNA assay, and the output module activity features are returned as the Net assay in the same Seurat object. Users may conveniently conduct dimension reduction and cell-clustering based on the Net assay, and perform marker genes analysis based on the RNA assay, by switching the default assay of Seurat object.

Construction of Cell State-Specific Characteristic Molecular Interaction Network (CMIN)

To annotate a certain cell state, beyond the marker genes, **SCORE** constructs the CMIN with the concept of Steiner Tree in graph theory. Given a graph $G = (V, E)$ and a subset of vertices $T \subset V$ (called terminal vertices), a Steiner tree $S \subset G$ is a connected tree that spans through the given terminal vertices T. The Steiner tree S may contain vertices not presented in T, known as the Steiner vertices, serving as the interchange node to connect the vertices in T. In molecular interaction network, the

marker genes of a certain cell state are typically selected as the terminal vertices, and the mediating Steiner vertices, although not necessarily differentially expressed, are supposed to play important roles in formulating the specific cell state through molecular interactions. Therefore, the Steiner tree provides a simplified representation of original PPI network, and highlights the significance of non-marker interacting genes surpassing traditional marker genes analysis.

In the downstream analysis of **SCORE**, given the specific cell cluster identified from module activation features, we first construct the terminal genes set T^* by detecting the marker genes from two different levels,

1) Union of all genes in the **SCORE**-extracted modules that are differentially activated in the cluster, denoted as differentially activated module genes (DAMGs);

2) Individual genes that are significantly up-regulated in the cluster, denoted as differentially expressed genes (DEGs).

While the DEGs are commonly referred as the "markers" of cellular states, the DAMGs also represent the key molecular interaction modules to mark the cell cluster in the network resolution.

Next, to infer the CMINs that possibly formulate the cellular states rather than the genes solely mark the cellular states, we propose to calculate a Steiner tree S^* that spans the terminal gene set $T^*$with some optimal property, defined on the constructed WMIN $\mathcal{G}(V, E, W)$ by **SCORE** in the first step of workflow. We require that the separate DAMGs or DEGs in S^* are linked by the most relevant genes, as well as through the most likely interaction path derived from the dataset. To this end, we define the distance D_{ij} on the edge E_{ij} of WMIN by $D_{ij} = 1/(W_{ij} + \varepsilon)$, where the small number ε is added to avoid zero in dominator. Highly correlated genes pairs in PPI tend to possess much closer distance from the definition. Then S^* can be optimized as the Steiner tree with the least sum of edge distances, which can be tackled efficiently by the greedy algorithm.

For the better visualization of CMIN, in R implementation of **SCORE** we mark DAMGs, DEGs and Steiner connecting genes with different colors, and also use the PageRank algorithm to measure the topological importance of the genes in optimal Steiner tree S^* as shown by the size of the nodes.

Settings in the benchmarking gold-standard dataset

The gene expression matrix was processed as fragments per kilobase per million reads (FPKM) as in the original literature. To test the robustness of different methods, we first adopted the vst method in Seurat v3.0 package to select five groups of highly variable genes (with the number of genes 2000, 5000, 8000, 11000 and 14000 respectively). We performed SCORE, as well as two other network or biological information based methods, CSN and SCENIC to further compress and extract the features respectively from the groups of highly variable genes (HVGs).

For **SCORE**, we downloaded the *Homo sapiens* PPI network (version 3.5.173) from the BioGRID database. The top ranked genes included in the calculation of AUC values varied with the sizes of input HVGs, with 250 and 200 for 2,000 and 5,000 variable genes respectively, and 400 in other cases. The parameters in implementing CSN and SCENIC were chosen as default. The SCENIC yields both continuous and binary features as the outputs.

The running time comparison was conducted on the 2.50GHz Xeon E5-2680 machine with 128G RAM, 12 cores and Linux OS. For SCORE and SCENIC, the CPU core number was set as 10. CSN was automatically paralleled with MATLAB 2019b. The wall times of implementing each procedure were recorded as the running times in the main text.

In the downstream clustering analysis, three methods, SNN, SIMLR and SC3 were performed on the extracted features by different methods, and the adjusted rand index

(ARI) as well as the similarity matrix of SC3 were used to evaluate the accuracy and the effect of batch removal. The direct analysis on raw expression matrix with selected HVGs was also performed for the comparison. The true labels were the seven collected cell lines identity (H1, GM12878, A549, HCT116, H1437, K562 and IMR90) without batch information. For SNN, we tuned the resolution parameter to obtain the optimal ARI value. As to SC3 and SIMLR, we set the number of clusters to 7. The t-SNE plot was produced on the top 10 principal components of the extracted features.

Human embryonic datasets

To evaluate the integration performance of SCORE, five human embryonic datasets were collected from our previously published studies, including fetal germ cells (FGCs) (GEO number: GSE86146), heart (GEO number: GSE106118), kidney (GEO number: GSE109488), prefrontal cortex (PFC) (GEO number: GSE104276), and cerebral cortex (GEO number: GSE103723), spanning from 4 to 26 weeks of fetal development. Importantly, we re-organized the five datasets using uniform pipeline and format, which provided a rich and convenient resource for studying human embryonic development (https://github.com/zorrodong/HECA). In brief, barcode and UMI information were extracted by UMI-tools form raw reads[29]. After discarding the poly A bases, TSO sequences and low-quality sequences, the clean reads were mapped to GRCh38 reference using STAR aligner[30]. We used featureCounts[31] to annotate the mapped reads and quantified the UMI counts through UMI-tools. We provided the pipeline for users (https://github.com/zorrodong/HECA).

To analyze the human embryonic datasets, we first discard cells with gene number below 1,000 and UMI counts below 10,000. HVGs were chosen using Seurat (mean >= 0.1, dispersion >= 0.1), and about 8,000 HVGs were selected for SCORE to perform the evaluation. Two batch effect correction methods CCA and Harmony were used with the recommended parameters. 30-50 reduced dimensions were used to perform UMAP analysis and clustering using the graph-based method in Seurat. To accelerate the speed, we used genesorteR to conduct the differential expression

analysis.

Human adult small intestine dataset

This study was supported by the Ethics Committee of Peking University Third Hospital (License No. M2016170). All patients had signed written informed consent. Tissues were collected from 6 small intestine cancer patients immediately after surgical resection. The adjacent normal tissues were sampled at least 5 cm away from cancer border. We dissociate the samples into single cells and constructed the libraries using a modified STRT technique following the previous protocol[32]. Libraries were sequenced using Illumina HiSeq 4000 platform with 150-bp paired-end reads. We also used the uniform pipeline as described above to analyze the human adult small intestine dataset.

Data availability

Human adult small intestine dataset are deposited in the GEO under the accession number: XXXXXX.

Acknowledgements

This project was supported by grants from the National Natural Science Foundation of China (31625018 and 81521002 to F.T.).

Author contributions

F.T., T.L., W.F., J.D. conceived the project; W.W., X.Z., L.W., performed the experiments; J.D., P.Z., Y.W., Y.C., H.X., J.L., X.N.Z. conducted the bioinformatics

analyses; J.D., P.Z., F.T., wrote the manuscript with the help of all the authors.

Competing interests

The authors declare no competing financial interests.

References:

1 Arendt, D. *et al.* The origin and evolution of cell types. *Nature Reviews Genetics* **17** (2016).

2 Achim, K. & Arendt, D. Structural evolution of cell types by step-wise assembly of cellular modules. *Current Opinion in Genetics & Development* **27**, 102-108 (2014).

3 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377-382 (2009).

4 Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187-1201 (2015).

5 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

6 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).

7 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15, doi:10.1186/s13059-017-1382-0 (2018).

8 Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput Biol* **11**, e1004575, doi:10.1371/journal.pcbi.1004575 (2015).

9 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086, doi:10.1038/nmeth.4463 (2017).

10 Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods* **13**, 241-+, doi:10.1038/Nmeth.3734 (2016).

11 Dai, H., Li, L., Zeng, T. & Chen, L. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res* **47**, e62, doi:10.1093/nar/gkz172 (2019).

12 Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273-282, doi:10.1038/s41576-018-0088-9 (2019).

13 Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47**, D529-D541, doi:10.1093/nar/gky1079 (2019).

14 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).

15 Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8**, 15599, doi:10.1038/ncomms15599 (2017).

16 Ronen, J. & Akalin, A. netSmooth: Network-smoothing based imputation for single cell

RNA-seq. *F1000Research* **7**, 8, doi:10.12688/f1000research.13511.3 (2018).

17 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics* **49**, 708-718, doi:10.1038/ng.3818 (2017).

18 Xu, C. & Su, Z. C. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974-1980, doi:10.1093/bioinformatics/btv088 (2015).

19 Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14**, 414-416, doi:10.1038/nmeth.4207 (2017).

20 Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486, doi:10.1038/nmeth.4236 (2017).

21 Li, L. *et al.* Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions. *Cell stem cell* **20**, 891-892, doi:10.1016/j.stem.2017.05.009 (2017).

22 Cui, Y. *et al.* Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell reports* **26**, 1934-1950 e1935, doi:10.1016/j.celrep.2019.01.079 (2019).

23 Wang, P. *et al.* Dissecting the Global Dynamic Molecular Profiles of Human Fetal Kidney Development by Single-Cell RNA Sequencing. *Cell reports* **24**, 3554-3567 e3553, doi:10.1016/j.celrep.2018.08.056 (2018).

24 Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528, doi:10.1038/nature25980 (2018).

25 Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell research* **28**, 730-745, doi:10.1038/s41422-018-0053-3 (2018).

26 Korsunsky, I. *et al.* Fast, sensitive, and accurate integration of single cell data with Harmony. *bioRxiv*, 461954, doi:10.1101/461954 (2018).

27 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420, doi:10.1038/nbt.4096 (2018).

28 Wang, N. & Teschendorff, A. E. Leveraging high-powered RNA-Seq datasets to improve inference of regulatory activity in single-cell RNA-Seq data. *bioRxiv*, 553040, doi:10.1101/553040 (2019).

29 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

30 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

31 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

32 Dong, J. *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome biology* **19**, 31, doi:10.1186/s13059-018-1416-2 (2018).

Figure 1

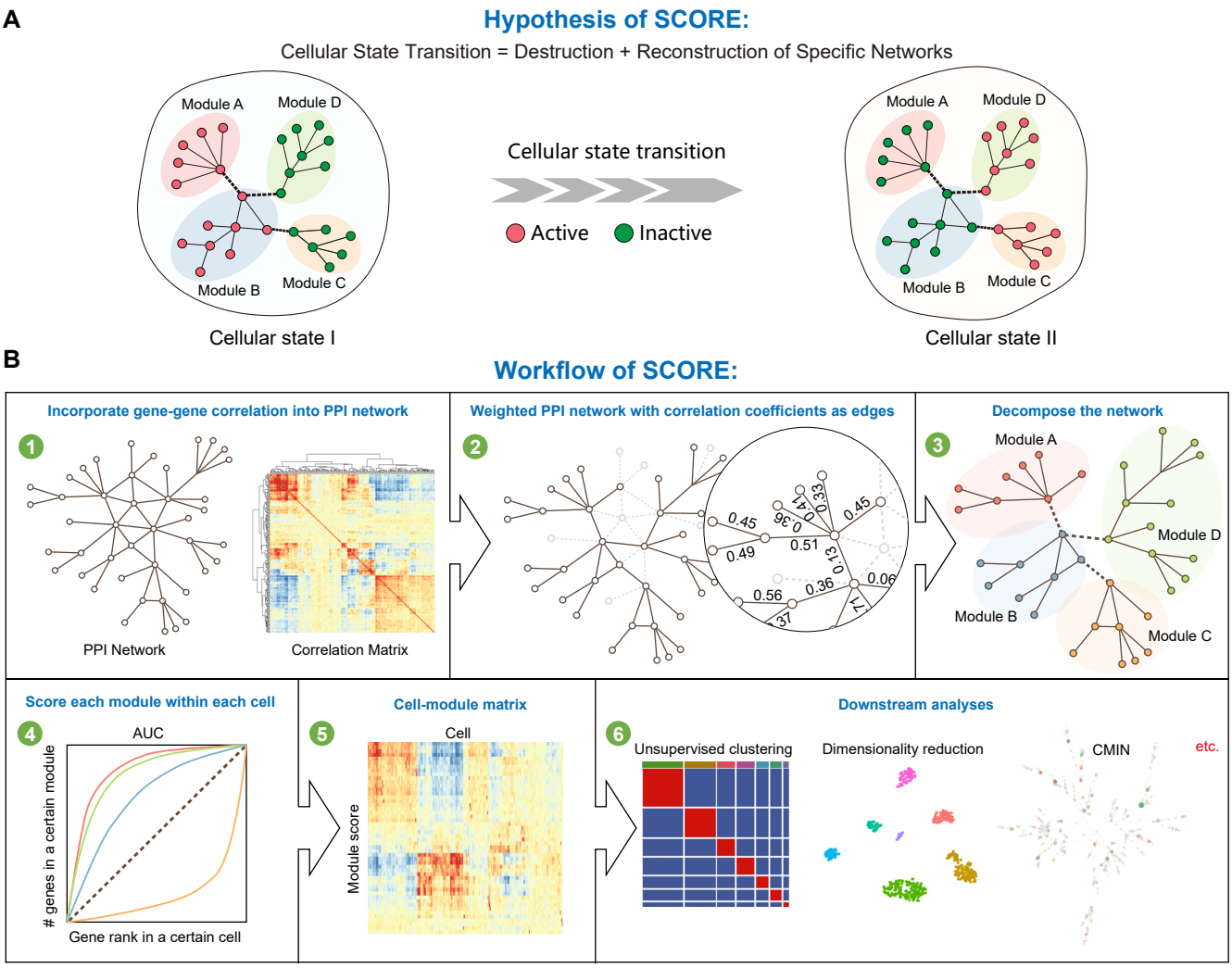

A
Hypothesis of SCORE:
Cellular State Transition = Destruction + Reconstruction of Specific Networks
Module A
Module D
Module B
Module C
Cellular state I
Cellular state transition
Active
Inactive
Module A
Module D
Module B
Module C
Cellular state II
B
Workflow of SCORE:
Incorporate gene-gene correlation into PPI network
1
PPI Network
Correlation Matrix
Weighted PPI network with correlation coefficients as edges
2
0.45
0.49
0.36
0.41
0.33
0.45
0.51
0.13
0.56
0.36
0.06
37
71
Decompose the network
3
Module A
Module D
Module B
Module C
Score each module within each cell
4
AUC
genes in a certain module
Gene rank in a certain cell
Cell-module matrix
5
Cell
Module score
Downstream analyses
6
Unsupervised clustering
Dimensionality reduction
CMIN
etc.

Figure 2

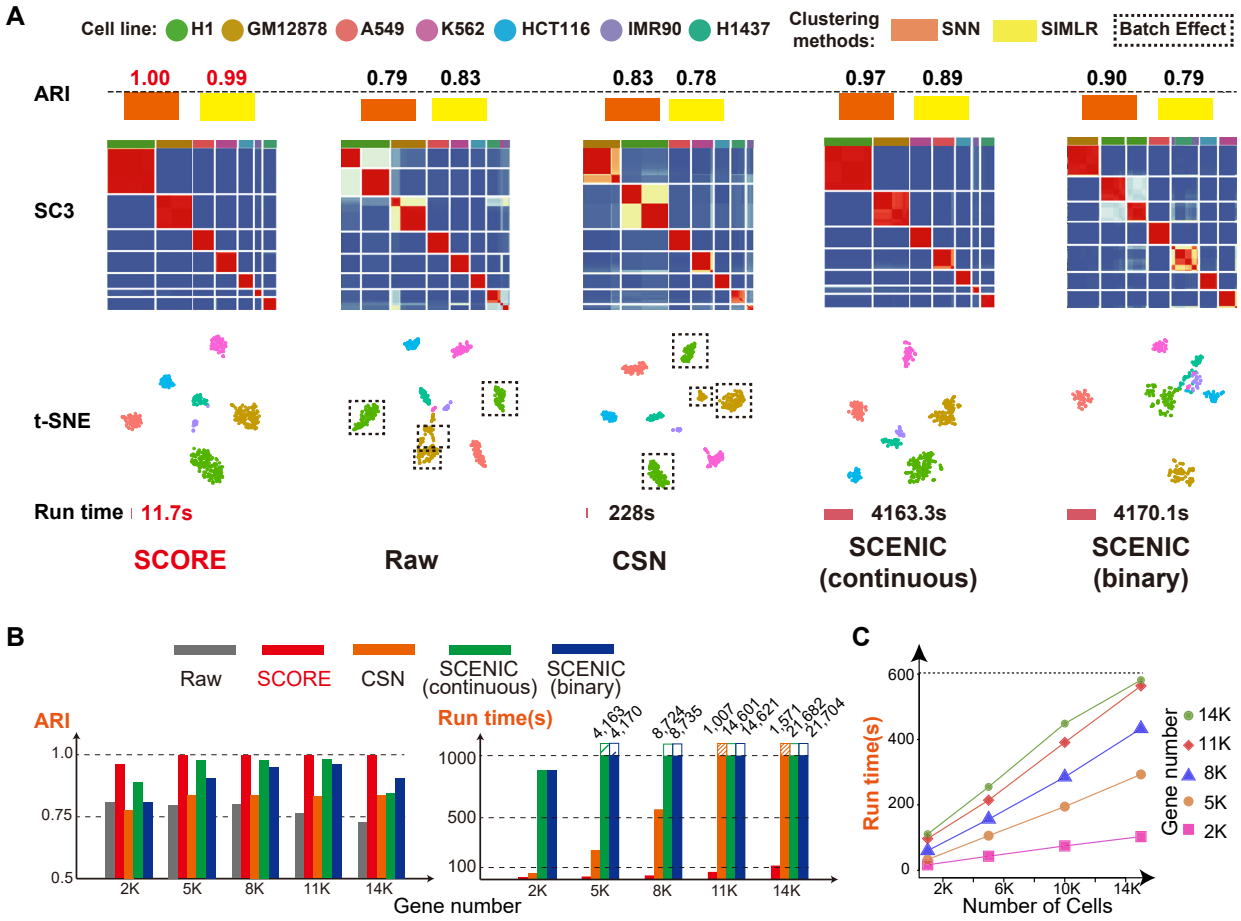

A
Cell line: H1 GM12878 A549 K562 HCT116 IMR90 H1437
Clustering methods: SNN SIMLR Batch Effect
ARI
1.00 0.99 0.79 0.83 0.83 0.78 0.97 0.89 0.90 0.79
SC3
t-SNE
Run time 11.7s 228s 4163.3s 4170.1s
SCORE
Raw
CSN
SCENIC (continuous)
SCENIC (binary)
B
Raw SCORE CSN SCENIC (continuous) SCENIC (binary)
ARI
1.0
0.75
0.5
2K 5K 8K 11K 14K
Run time(s)
1000
500
100
4,163 4,170 8,724 8,735 1,007 14,601 14,621 1,571 21,682 21,704
Gene number
C
Run time(s)
600
400
200
0
2K 6K 10K 14K
Number of Cells
Gene number
14K
11K
8K
5K
2K

Figure 3

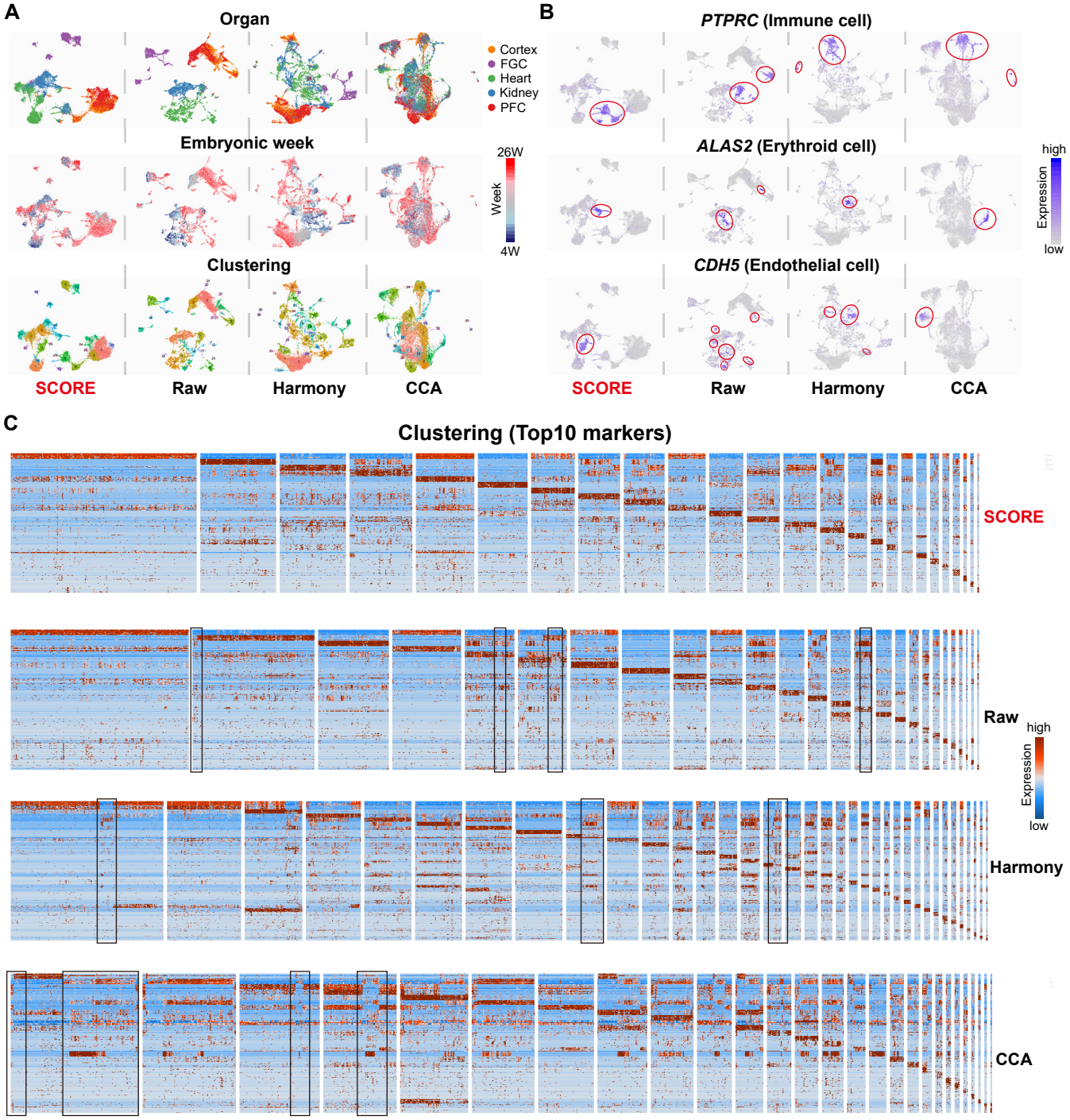
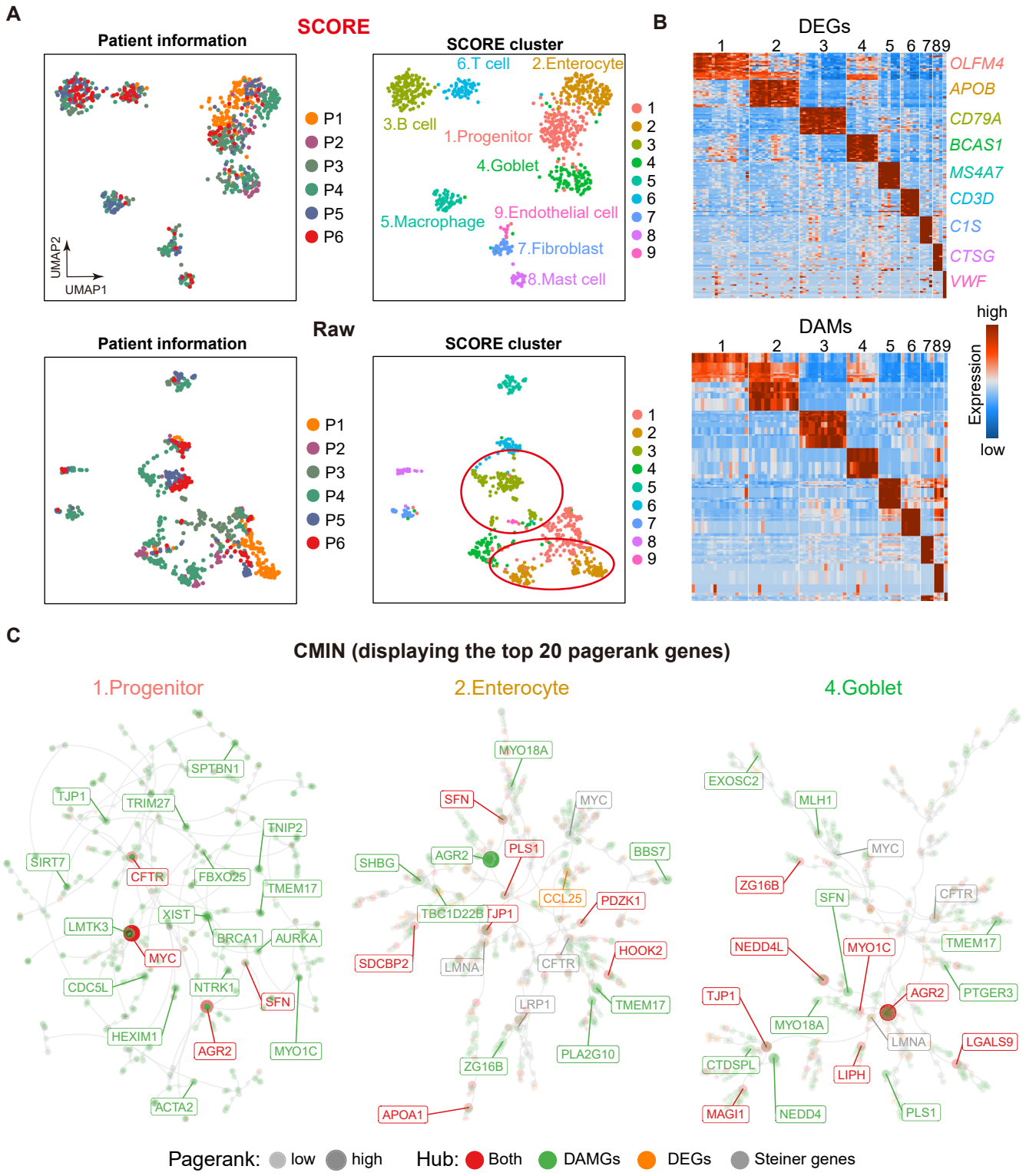

A
Organ
Cortex
FGC
Heart
Kidney
PFC
Embryonic week
26W
Week
4W
Clustering
SCORE
Raw
Harmony
CCA
B
PTPRC (Immune cell)
ALAS2 (Erythroid cell)
CDH5 (Endothelial cell)
high
Expression
low
SCORE
Raw
Harmony
CCA
C
Clustering (Top10 markers)
SCORE
Raw
Harmony
CCA
high
Expression
low

Figure 3


A
SCORE
Patient information
SCORE cluster
6.T cell
2.Enterocyte
3.B cell
1.Progenitor
4.Goblet
5.Macrophage
9.Endothelial cell
7.Fibroblast
8.Mast cell
P1
P2
P3
P4
P5
P6
1
2
3
4
5
6
7
8
9
UMAP2
UMAP1
B
DEGs
OLFM4
APOB
CD79A
BCAS1
MS4A7
CD3D
C1S
CTSG
VWF
Raw
DAMs
high
low
Expression
C
CMIN (displaying the top 20 pagerank genes)
1.Progenitor
2.Enterocyte
4.Goblet
SPTBN1
TJP1
TRIM27
TNIP2
SIRT7
CFTR
FBXO25
TMEM17
XIST
LMTK3
BRCA1
AURKA
MYC
CDC5L
NTRK1
SFN
HEXIM1
AGR2
MYO1C
ACTA2
MYO18A
SFN
MYC
SHBG
AGR2
PLS1
BBS7
CCL25
PDZK1
TBC1D22B
JP1
SDCBP2
LMNA
CFTR
HOOK2
LRP1
TMEM17
ZG16B
PLA2G10
APOA1
EXOSC2
MLH1
MYC
ZG16B
SFN
CFTR
NEDD4L
MYO1C
TMEM17
TJP1
AGR2
PTGER3
MYO18A
LMNA
LGALS9
CTDSPL
LIPH
MAGI1
NEDD4
PLS1
Pagerank:
low
high
Hub:
Both
DAMGs
DEGs
Steiner genes