JOURNAL OF XXX, VOL. 14, NO. 8, MARCH 2020

TiC2D: trajectory inference from single-cell RNA-seq data using consensus clustering

Yanglan Gan, Ning Li, Cheng Guo, Guobing Zou, Jihong Guan* and Shuigeng Zhou*, *Member, IEEE*

Abstract—Cellular programs often exhibit strong heterogeneity and asynchrony in the timing of program execution. Single-cell RNA-seq technology has provided an unprecedented opportunity for characterizing these cellular processes by simultaneously quantifying many parameters at single-cell resolution. Robust trajectory inference is a critical step in the analysis of dynamic temporal gene expression, which can shed light on the mechanisms of normal development and diseases. Here, we present TiC2D, a novel algorithm for cell trajectory inference from single-cell RNA-seq data, which adopts a consensus clustering strategy to precisely cluster cells. To evaluate the power of TiC2D, we compare it with three state-of-the-art methods on four independent single-cell RNA-seq datasets. The results show that TiC2D can accurately infer developmental trajectories from single-cell transcriptome. Furthermore, the reconstructed trajectories enable us to identify key genes involved in cell fate determination and to obtain new insights about their roles at different developmental stages.

Index Terms—Trajectory inference; Consensus clustering; Gene partition; Single-cell transcriptome.

1 Introduction

Cellular programs are governed by inherent stochasticity and a vast external factors [1]. During the process of program execution, cells undergo a complex choreography of gene regulatory changes, often exhibiting strong heterogeneity and asynchrony [2]. Understanding cell state transitions can provide new perspectives into complex cellular dynamic processes, such as cell differentiation, development or diseases [3]. Recent advances in single-cell RNA-sequencing (scRNA-seq) open up new frontiers of studying the fundamental mechanisms underlying normal development and diseases [4, 5]. Based on massive single-cell gene expression data, such dynamic processes can be modeled computationally using trajectory inference methods, which order cells by progression and reconstruct the pseudo-time trajectories as they undergo some types of biological transitions [6].

To infer the trajectories of cell state transitions, current computational approaches broadly fall into two categories [7]. The predominant category of algorithms directly estimates the pseudo-time ordering from cell profiles. Monocle [8] describes the complex transition structure using a minimum spanning tree (MST), which serves as the pseudo-time axis to place cells in order. Wanderlust [9] infers a linear trajectory based on the k-nearest neighbor graph. Wishbone [10] extends the Wanderlust method to position individual cells along bifurcating developmental trajectories. DPT [11] applies a diffusion map and random-walk-based distance to directly reconstruct cell differentiation trajectories. STREAM [12] projects cells to a lower dimensional space using MLLE embedding and infers trajectories using the Elastic Principal Graph. In these methods, the high-dimensional transcriptional profiles are usually projected into reduced dimensional spaces by using a (non)linear transformation. The other category of solutions is to take advantage of clustering in trajectory inference. These methods are based on the assumption that the high-dimensional cellular profiles constitute cell subpopulations at distinct stages, and therefore various clustering methods can be utilized to identify these cell subpopulations. Subsequently, the identified clusters are served as anchor points, which facilitate the process of trajectory inference. TSCAN [13] proposes a cluster based minimum spanning tree approach to order cells. The process of clustering the cells could effectively reduce the complexity of the tree space. Mpath [14] first clusters the cells, then designates landmark clusters with each representing discrete cell states, and finally derives multi-branching trajectories using neighborhood-based cell state transitions. Monocle 2 [15] is also based on the construction of a minimum spanning tree, with an intermediate clustering step. SLICE [16] quantitatively measures the cellular differentiation states based on the calculated single-cell entropy and further infers a lineage model by constructing a directed minimum spanning tree among the identified clusters. PAGA [17] provides an interpretable graph-like map of the data manifold, by estimating the connectivity of manifold partitions. GraphDDP [7] starts from a user-defined cluster assignment and then utilizes a force-based graph layout approach to detect differentiation pathways.

These existing methods have made significant progress in trajectory inference. However, due to noise, high dimensionality and data heterogeneity, most of them still suffer from some limitations. Firstly, dimensionality reduction techniques usually limit their ability to recover the complex

- *Y.Gan, N.Li and C.Guo are with the School of Computer Science and Technology, Donghua University, Shanghai, China.*
- *G.Zou is with the School of Computer Engineering and Science, Shanghai University, Shanghai, China.*
- *J.Guan is with the Department of Computer Science and Technology, Tongji University, Shanghai, China.*
- *S.Zhou is with the Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China.*

Manuscript received XXXX; revised XXXX.

structures of data [18]. In particular, principal component analysis (PCA) explains only a small fraction of data variance and hence does not offer a clear separation for all factors [19]. Constrained by the global arrangement, multi dimensional scaling (MDS) easily ends up distorting the local arrangement [20]. The widely used t-SNE depends on a perplexity parameter, which is hard to be determined. Incorrect values may lead to a layout where data points divided into several detached groups are arbitrarily positioned [21]. Secondly, as the clustering based methods usually adopt traditional clustering algorithms, such as k-means, hierarchical clustering and spectral clustering, they are sensitive to noise and high dimensionality [22]. Recently, consensus clustering has been proposed for RNA-seq data clustering. The method ECC adopts consensus clustering strategy for patient stratification [23]. SC3 is proposed to cluster cells by consensus clustering [24]. These methods achieve high accuracy and robustness by combining multiple clustering solutions through a consensus approach. Thus, it implies a feasible way to develop more robust trajectory inference methods, and to fuse diverse solutions from multiple heterogeneous datasets.

In this study, we present a new method termed *TiC2D* for reconstructing cell developmental trajectories from single-cell RNA-seq data, by introducing a consensus clustering model to precisely cluster cells and infer cell trajectory based on the identified clusters. To uncover interesting clusters from data with noise, outliers and data variation, TiC2D first adopts a community detection strategy to separate genes into different groups, and clusters the cells into different partitions for each gene group, then ensembles these partitions into consensus clusters. Subsequently, an MST is constructed to connect cluster centers. Furthermore, the principal curves algorithm is utilized on the MST to get smooth trajectories. To evaluate the performance of TiC2D, we applied it to four independent scRNA-seq datasets with known lineages and developmental time information. Our analyses show that the proposed TiC2D algorithm can successfully reconstruct cellular trajectories that have been previously experimentally validated. Meanwhile, the comparison with state-of-the-art methods indicates that consensus clustering significantly helps the generation of more robust trajectories.

2 Materials and Methods

2.1 Datasets and data prepossessing

To demonstrate the effectiveness of TiC2D, we applied it to four real scRNA-seq datasets from both Homo sapiens and Mus musculus. The first dataset consists of single-cell RNA-seq samples collected after stimulating bone-marrow-derived mouse dendritic cells by lipopolysaccharide (LPS) [25]. It contains a total of 306 cells collected at 1, 2, 4 and 6 h. The second dataset consists of 372 primary human skeletal muscle myoblasts (HSMM) single cells collected at 0, 24, 48, and 72 h [8]. The third dataset is germline-human-male-weeks datasets(Germline), and contains 649 cells collected at 4, 9, 10, 19, 20, 21 and 25 weeks [26]. The fourth dataset E-MTAB-3929 contains 1,529 cells obtained from 88 human preimplantation embryos ranging from embryonic day 3 to 7 [27]. The original quantification and cell types of these datasets are described in the related articles. In the following, we respectively refer to these four datasets as LPS, HSMM, Germline and MTAB.

For the four datasets, we preprocess the scRNA-seq dataset as commonly done in previous studies [28], including basic filtering of the data, normalization and log2-transformation. In order to alleviate the effect of drop-out events on the following analyses, the genes with zero read count in all cells are firstly filtered out [13]. Then the gene expression values are normalized, and the resulted values are further log2 transformed by adding a pseudo-count of 1. As the number of remaining genes is still much large, we adopt the gene filter to remove genes that are either expressed in less than X% of cells or expressed in at least (100 – X)% of cells, as the previous method SC3 did (Kiselev et al., 2017). To select an appropriate value of the filtering parameter X, we try different values of X ranging from 5% to 20% with the step 0.025, and find that in general the gene filter does not greatly affect the accuracy of clustering (Supplementary Table 1). However, the gene filter significantly reduces the dimensionality of the data.

2.2 The TiC2D Method

To reconstruct trajectories from a heterogeneous population of single-cells, we proposed a new trajectory inference method TiC2D. The method consists of four main steps, as shown in Figure 1. First, divide the genes into different groups using a community detection strategy. Second, for each gene group, respectively cluster cells into different subpopulations. Third, ensemble the different partitions of these gene groups into consensus clusters. Finally, construct a minimum spanning tree to connect the identified consensus clusters, and then get smooth trajectories by principal curves algorithm [19].

Step 1. Partitioning genes into different groups

We treated the preprocessed gene expression data after gene filtering as a matrix E_{M*N} with M genes expressed in N cells. Element E_{ij} represents the expression level of the ith gene in the jth cell. The filtered gene set was partitioned into different subgroups using a community detection algorithm. Concretely, we first constructed a kNN graph and identified gene subgroups by detecting communities of densely interconnected genes. Here, we applied the method Louvain for community discovery, which is an accurate and efficient algorithm based on multi-level optimization of modularity [29]. Subsequently, E_{M*N} is partitioned into K submatrices E_1, E_2, $\cdots$, E_K. E_k represent the gene expression submatrix of N cells in the kth gene group.

Step 2. Clustering cells for each gene group

For each gene group E_k, we clustered N cells into R_k basic partitions using a clustering method (e.g. Louvain, k-means). Based on the basic partitions, a binary similarity matrix is constructed according to the corresponding cluster labels of cells. The binary matrix is represented as B_{N*R_k}. For each row, only one element is 1, the others are 0, representing that this cell only belongs to this cluster. Totally, for all the K gene subgroups, we obtained K basic partitions for these N cells and correspondingly K binary matrices $B_{N*R_1}, B_{N*R_2}, \cdots, B_{N*R_K}$.

Step 3. Ensembling clusters from different cluster sets

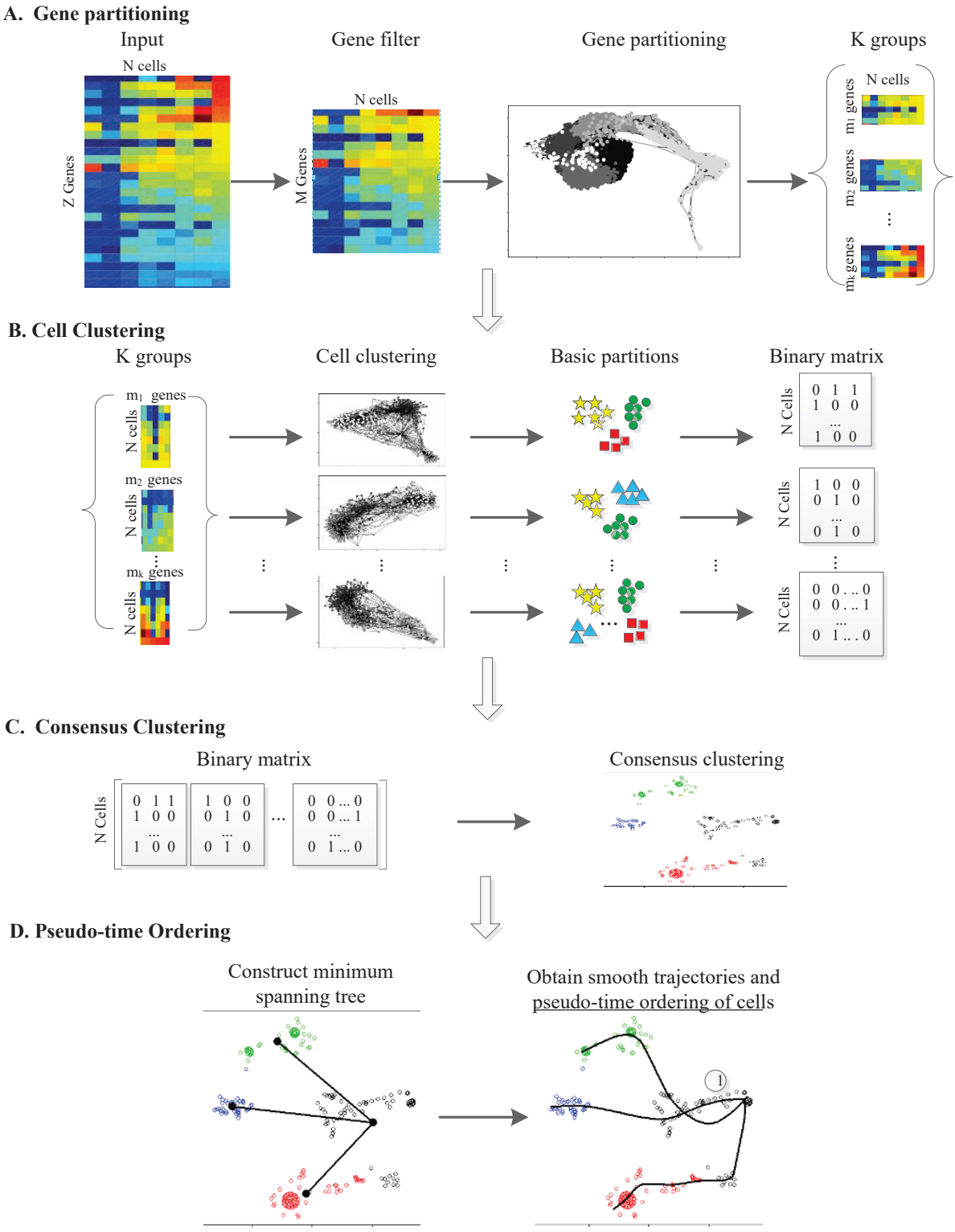


Fig. 1. The schematic flowchart of TiC2D. (A) Filtering genes and partitioning genes into different groups. (B) For each gene group, clustering *N* cells into different subpopulations. (C) Ensembling different partitions of these gene groups into consensus clusters. (D) Obtaining the pseudo-time ordering based on consensus clusters .

As traditional clustering algorithms easily suffer from noise and data heterogeneity associated with high-throughput molecular data, we adopted consensus clustering to merge these independently generated basic partitions, and ensured that the final consensus partitions maximally agree with the basic ones. With the K different partitions, we concatenated all those binary matrices to form a larger binary matrix $B = \{B_{N*R_k} — k=1, 2, \cdots, K\}$. Furthermore, we performed k-means clustering on the merged binary matrix B to obtain consensus clusters.

Step 4. Pseudo-time ordering

Based on the consensus clusters, a minimum spanning tree with the smallest total edge length is constructed to connect all the cluster centers. The minimum spanning tree determines the initial structure of the trajectory. To further infer the origin of the trajectory, we calculated the average similarity of the *m*th cluster, denoted by aSC_m, is computed as follows:

$$aSC_m = \frac{\sum_{i=1}^{Cm}\sum_{j=1,j\neq i}^{Cm} Si,j}{Cm(Cm-1)} \quad (1)$$

$S_{i,j}$ denotes the similarity between two cells *i* and *j*, and C_m is the number of cells in the *m*th cluster. The similarity can be defined as cosine similarity, Pearson correlation coefficient or Jaccard coefficient, etc. This metric evaluates the average similarity among the cells in the cluster. As previous studies have shown that differentiation pathways follow the gradient of similarity from higher progenitors towards lower mature cell types [7], the cluster with the highest average similarity may be regarded as the starting point of the differentiation path.

Next, the individual cells are further projected onto tree edges to infer the cell-level trajectory along the main path and each branches. The pseudo-time is constructed separately for the main path and each branches by using the principal curves algorithm (implemented in the princurve R package) [19]. For each path, the pseudo-time ordering is determined by projecting all related data points onto the curve and calculating the distance from the starting point of the curve to each projected data point. Then this algorithm smoothens the trajectory iteratively until convergence.

2.3 Evaluation metrics

We adopted the adjusted Rand index (ARI) to measure the accuracy of clustering methods [30]. For a set of *N* cells and two different partitions of these cells, the overlap between the two partitions can be summarized in a contingency table, in which each entry denotes the number of objects in common between the two partitions. The ARI is then calculated as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_i}{2}]/\binom{n}{2}}{[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}]/2 - [\sum_i \binom{a_i}{2}\sum_j \binom{b_i}{2}]/\binom{n}{2}} \quad (2)$$

where (.) denotes a binomial coefficient, $n_{i,j}$ is the element from the contingency table, a_i is the sum of the *i*th row of the contingency table, b_j is the sum of the *j*th column of the contingency table. The range of ARI is [-1,1]. The larger the value, the more consistent the clustering result is with the reference one.

For a trajectory inference algorithm, the cell ordering accuracy is evaluated based on Pseudo-temporal Ordering Score (POS) [13] and Kendall's tau correlation [12, 15]. It is assumed that external information not used in pseudo-time reconstruction is available to evaluate the pairwise order of cells. Its main idea is characterizing how well the ordering of cells in the inferred trajectory matches the true one.

The POS of a trajectory is defined as:

$$POS = \sum_{i=1}^{N-1}\sum_{j:j>i} g(i,j) \quad (3)$$

$$g(i,j) = \begin{cases} 0 & \text{if } T(i) = T(j) \\ (T(j) - T(i))/D & \text{otherwise} \end{cases} \quad (4)$$

where N is the number of cells, the *i*th and *j*th cells in the inferred trajectory are respectively collected from time points $T(i)$ and $T(j)$, and D is chosen to normalize POS so that $POS \in [-1, 1]$. If two cells are collected at the same time point, $g(i,j)$=0; otherwise, $g(i,j)$ is positive if $T(i)$ represents an earlier time point, or negative if $T(i)$ represents a time point later than $T(j)$. $POS = 1$ represents that the inferred order of cells perfectly matches the order determined by the cells collection time. $POS = -1$ means that the inferred order is in the opposite direction of the true situation. A higher absolute value of POS score indicates that the algorithm can more accurately infer the pseudo-time trajectory.

3 Results

3.1 Trajectory inference and pseudo-temporal ordering

To evaluate the effectiveness of TiC2D, we performed extensive analyses on four widely used real scRNA-seq datasets (LPS, HSMM, Germline and MTAB), which were produced by different techniques from a variety of contexts in human and mouse. In order to reconstruct reliable pseudo-time trajectories, the preliminary requirement is to correctly identify cell clusters, serving as anchor points of the paths. Different from traditional methods, we proposed a consensus clustering approach to identify more reliable clusters. The gene sets are firstly partitioned into different subgroups using a community detection algorithm. For each gene group, we clustered the cells to obtain basic partitions, and then fused the basic partitions into consensus clusters. To evaluate the performance of the proposed method, we compared it with traditional clustering methods (PCA+K-Means and PCA+Louvain). The result indicates that the performance of the consensus clustering is superior on these four real datasets (Figure 2), which imply that the consensus clustering helps us reduce the effect of noise and high dimensionality, which ensures the stability and accuracy of the final results. As shown in Figure 3(A), TiC2D respectively identifies 4, 4, 7, 5 clusters with the highest ARI scores from the datasets LPS, HSMM, Germline and MTAB.

With the identified clusters, we calculated the average pairwise similarity between all cells in each cluster to measure the average similarity of cluster. Figure 3(B) shows the aSC_m of the identified consensus clusters in the four datasets. By investigating the cells in these clusters, we noticed that the denser clusters usually represent progenitor cells, whereas the mature cell types exhibit lower similarity. For the dataset HSMM [8], this cell population comprises single-cell RNA-seq samples from differentiating human skeletal muscle myoblasts. TiC2D first identifies four clusters, which mainly correspond to the cell groups collected at 0, 24, 48 and 72 h. Correspondingly, the average similarities of these clusters are 0.747, 0.735,0.722 and 0.440. Meanwhile,

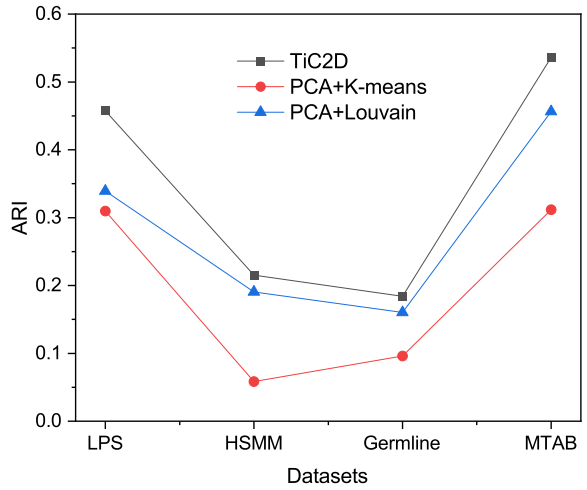


Fig. 2. The performance comparison among TiC2D, PCA+K-means and PCA+Louvain on four real scRNA-seq datasets (HSMM, LPS, Germline and MTAB).

confluent differentiation would materialize in the connections among clusters of similar similarity. This trend is also obvious in the other datasets. In the dataset LPS, we also identified 4 clusters. These clusters are in agreement with the original four cell groups, collected at 1, 2, 4 and 6 h after stimulating bone-marrow-derived dendritic cells by lipopolysaccharide. For the dataset Germline, we identified 7 subpopulations, corresponding to a variety of cell types collected at different stages. On the whole, by analyzing the average similarities of these clusters, we observed that a decreasing similarity pattern exists in these clusters, implying that the progenitor cells on average are more similar than their mature descendant cells.

Subsequently, a minimum spanning tree is constructed to determine the structure of trajectory. Figure 3(C) shows the reconstructed trajectories corresponding to these four datasets. We took the cluster with the highest average similarity as the starting point. Specifically, for the dataset HSMM, TiC2D obtains only a single main path. In the previous results reported by Monocle and TSCAN, a main path and a branch path have been inferred. However, the branching path in the minimum spanning tree constructed by Monocle and TSCAN was caused by contaminated interstitial mesenchymal cells [13]. Thus, the main path inferred by TiC2D is consistent with myoblast differentiation that is the biological process of interest. For the dataset MTAB, based on the five identified clusters, we also reconstructed a single path corresponding to the cell differentiation process from embryonic day 3 to 7. In the dataset LPS, cluster 1 has the highest average similarity, then we accordingly considered it as the origin of the pseudo-time trajectory. Cluster 2 and cluster 3 have similar average similarity, which respectively lie in two branches. Through constructing the minimum spanning tree from the identified consensus clusters, we gained two branching paths in the trajectory: ① → ② and ① → ③. For the dataset Germline, based on the constructed minimum spanning tree and the computed aSC, TiC2D identifies a trifurcating event in the pesudo-time trajectory: ① → ②, ① → ③ and ① → ④.

To evaluate the efficiency of these methods, we compared the execution time on different datasets. Specifically, to test whether TiC2D can scale well with larger scRNA-seq dataset, we applied it to simulated dataset generated by the Splatter simulator, which is used by previous studies [26, 31]. The simulated dataset contains a larger number of cells, consisting of 10000 cells and 4000 genes. The comparison result on the real and simulated datasets is shown in Supplementary Table 1. Among these methods, TSCAN has the highest efficiency. The efficiency of TiC2D is similar with Slingshot. As multiple steps of clustering on cells are carried out in our method, it consumes more time. However, since the individual clustering of cells are conducted on smaller partitions of gene set, which may lead to a comparable execution time.

3.2 Gene expression analysis along the inferred trajectories

Having demonstrated that TiC2D can accurately recapitulate the pseudo-times, we further analyzed to what extent a gene is involved in the trajectories. By ordering the gene expression with the pseudo-time, each part of the trajectory can be understood in terms of the behavior of certain genes. As in previous work [32], we adopted random forests to evaluate the importance of each gene with respect to the inferred trajectory. For each dataset, the genes are ranked by their importance, and the top ranked genes are further clustered into coherent gene modules. The number of clusters is automatically determined with the Bayesian information criterion [33].

To investigate the behavior of these important genes, we plotted the expression profiles of the top 100 genes of each dataset along the inferred pseudo-time trajectory. In Figure 4(A), the top ranked genes in the datasets LPS and Germline are respectively plotted. According to their expression levels in the cellular dynamic progress, these important genes are further divided into several modules. For the dataset LPS, we examined the expressions of the top ranked genes over time and found these genes exhibit three different expression patterns. In Module1 and Module2, the genes exhibit a gradually decreasing or increasing expression pattern, whereas the genes in Module3 show an approximately switch-like behaviour. To show more details, we displayed the expression of one representative gene along the pseudo-time trajectory for each module, as shown in Figure 4(B). For Module1, we can observe that the genes have high gene expression at the starting time while their expression levels decrease as the progression moves forward. On the contrary, in Module2, genes such as CCL22, TCF4 and NLRC5 are lowly expressed in the beginning, but are highly expressed in the post-progression. In Module 3, the expressions of genes (MS4A6D, SLAMF7, TRIM30A, etc) exhibit a switch-like pattern along the stimulation progression. For the dataset Germline, the top ranked genes are clustered into four modules. From the expression heatmap, we observed different expression patterns along the pseudo-time trajectory. For the datasets HSMM and MTAB, the expressions of the top ranked genes and their related modules are shown in the Supplementary Figure 1 and 2.

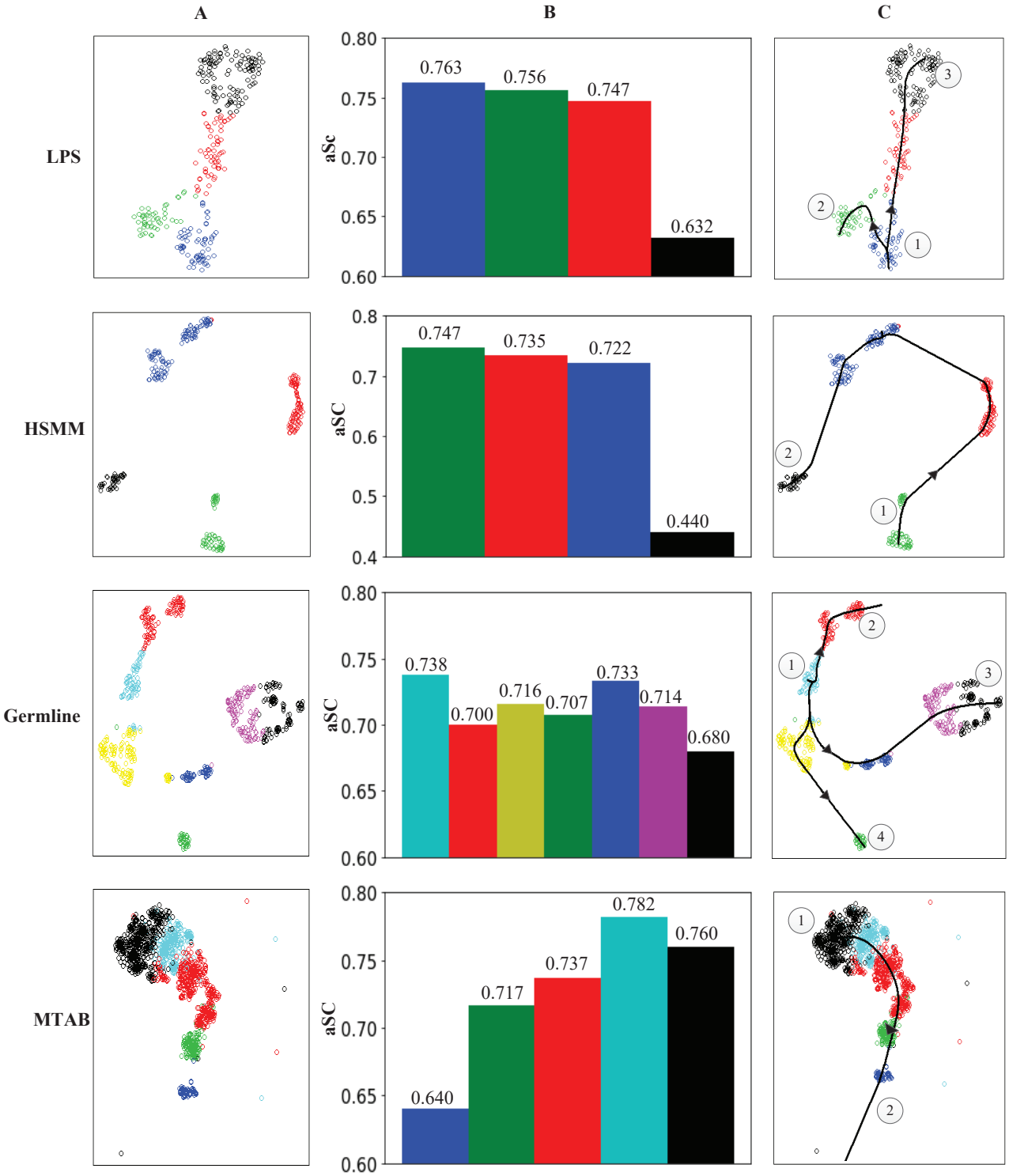


Fig. 3. TiC2D reconstructs the trajectory and pseudo-temporal ordering for four real scRNA-seq datasets (HSMM, LPS, Germline and MTAB). (A) Cell clusters identified by consensus clustering. Here t-SNE are used for the dimensionality reduction and display of these datasets. (B) The average similarities of identified clusters in each dataset. (C) The reconstruction of trajectory based on the identified clusters. The paths are reconstructed by concatenating pairwise shortest-paths between successive stable states in the minimum spanning tree of cells.① indicates the origin of each trajectory.

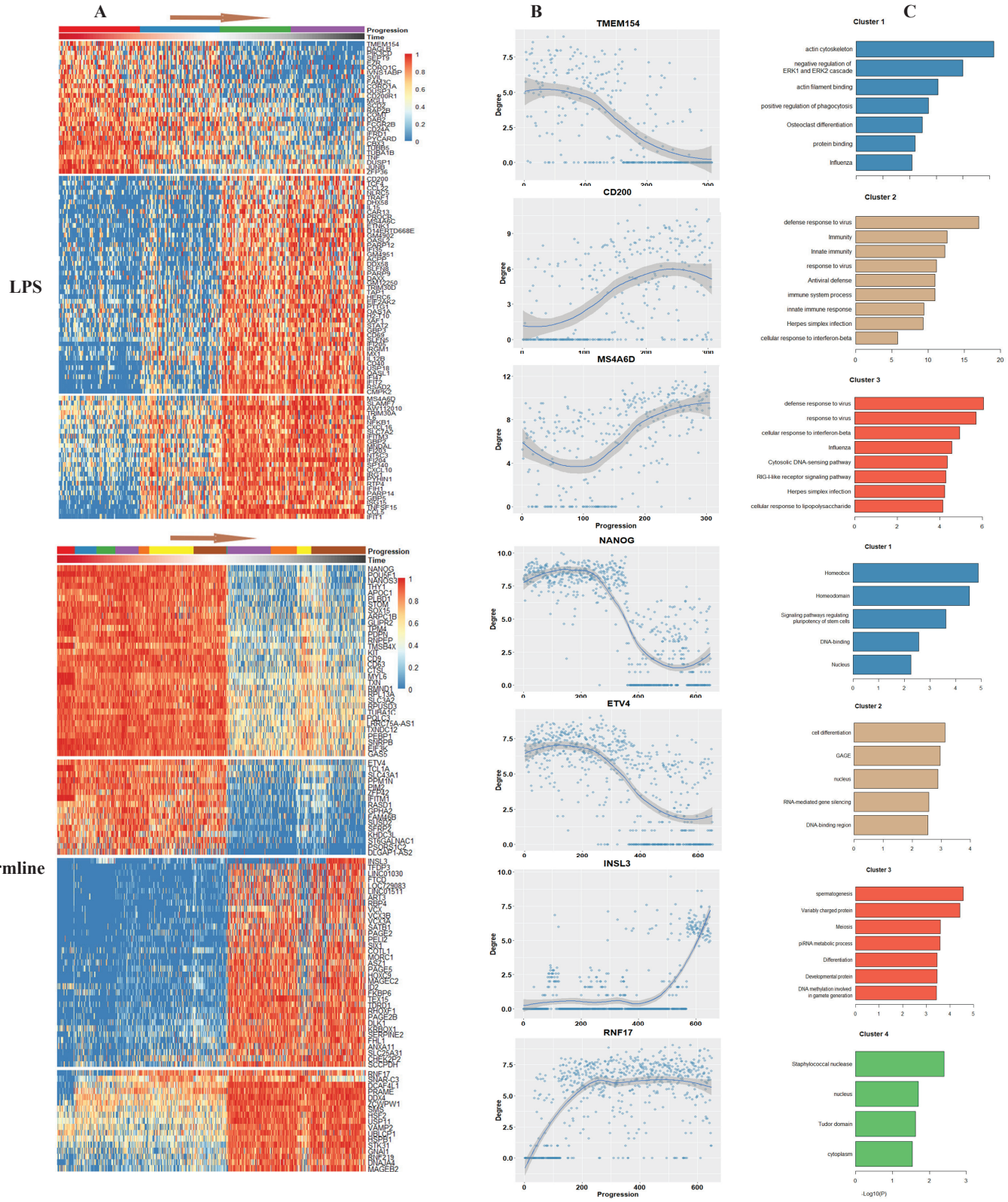


Fig. 4. The expression patterns and functional analysis of genes related to the inferred pseudo-time trajectories on two real scRNA-seq datasets (LPS and Germline). (A) The expression patterns of the top ranked genes. Progression means the pseudo-time cells order. (B) The detailed expression profiles of the representative gene in each module. The solid curves are plotted by the LOESS regression function. (C) Enriched functions of genes in each module.

Furthermore, we conducted gene function analysis to identify biological processes significantly enriched by the genes from each module (adjusted p value$<$0.05, DAVID gene ontology analysis). Based on the analysis, we observed that genes in each module not only exhibit coherent expression pattern, but also have very similar functions. For the dataset LPS, modules 2 and module 3 primarily contain genes that are involved in response to virus and immune system process (e.g. *OAS1A, OASL1, STAT2*) [34], as shown in Figure 4(C). The expression levels of these genes are relatively low in the progenitor cells, but rapidly increase after the stimulation. Specifically, OAS1A is known to be involved in the response to viral and bacterial stimulus, and its expression level is expected to increase after LPS stimulation. For the dataset Germline, the top ranked genes are grouped into four modules. In detail, Module 1 consists of genes that are involved in Homeobox and signaling pathways regulating pluripotency of stem cells, while genes in module 3 are mainly involved in cell differentiation and spermatogenesis (e.g. *FKBP6, MORC1*).

3.3 Performance comparison

We performed a comparative evaluation of TiC2D with state-of-the-art trajectory inference methods on four real scRNA-seq datasets. In the previous comparison [26], TSCAN, Monocle2 and Slingshot outperform most existing methods, and they are regarded as three excellent trajectory inference methods. Thus, we compared TiC2D with these three representative methods. Specifically, TSCAN determines the lineage structure by drawing an MST on the cluster centers [13]. Monocle2 constructs an explicit tree based on a scalable RGE algorithm DDRTree to order cells [15]. Slingshot also adopts a cluster-based MST to identify the global lineages and fits smooth branching curves to these lineages by simultaneous principal curves [31].

For each trajectory inference method, we ran ten times to get the average result. Figure 5(A)shows the performance comparison result based on POS. For all the four datasets, TiC2D has a higher POS than the compared three methods. For dataset LPS, the POS of TiC2D is about 0.87, while the highest POS of the other three methods is about 0.78. For dataset HSMM, TiC2D has a POS of 0.43, whereas the POS values of TSCAN and Monocle2 are just around 0.2. For dataset Germline, the POS of TiC2D is 0.41, only followed by the score of TSCAN (0.39). For dataset MTAB, Slingshot has similar performance with TiC2D. On the whole, TiC2D outperforms the three compared methods, indicating that the consensus clustering strategy and aSC_m measure are effective in reconstructing better pseudo-time. To further comprehensively compare the performance of these methods, we evaluate the inferred trajectories by the Kendall's tau correlation between the pseudotime assignment and the reference ordering (Qiu et al., 2017, Chen et al., 2019). The experiment results are shown in the Figure 5(B).

Subsequently, we compared the expression patterns of the marker genes along the true pseudo-time and the inferred orderings obtained by TiC2D, TSCAN, Monocle2 and Slingshot. The comparison results are shown in Figure 5. Generally, the gene expression profiles along the trajectories predicted by TiC2D exhibit much more similar patterns with the true pseudo-temporal trajectories than the other three methods. For dataset LPS, the expressions of the marker genes (CCL5, STAT2 and RSAD2) first increase and then slightly decrease along the true pseudo-time. The proposed TiC2D successfully reveals the increasing pattern of these genes (Figure 6(A)). By contrast, Monocle2 and Slingshot could not accurately reveal the temporal order, and the expression patterns of these genes are also less consistent with the true situation. For the dataset HSMM, the expression patterns of the marker genes (MEF2C, MYH3 and ENO3) along the trajectories predicted by TiC2D are more consistent with those along the true pseudo-order. For datasets Germline and MTAB, we observe similar trends. From a different perspective, these results validate that more accurate pseudo-time trajectories are inferred by TiC2D, which is consistent with the conclusion drawn from the results of POS.

4 Conclusion

While the advent of single-cell RNA-sequencing has shed new insights into cellular dynamic processes, it also raises new computational challenges. On the one hand, the expression data obtained by scRNA-seq is rather noisy, and thus computational models should take this factor into account. On the other hand, as in most cases no known markers exist to assign cells, it is hard to determine the starting points of the developmental pathways. In this paper, we proposed a new computational method TiC2D for inferring cellular trajectories and pseudo-time order from scRNA-seq data. To alleviate the effect of noise and high dimensionality, we adopted a consensus clustering model to precisely cluster cells on partitioned gene groups. Meanwhile, a measurement of cluster average similarity is applied to determining the origins of trajectories.

Using TiC2D, we reconstructed experimentally validated trajectories of four independent scRNA-seq datasets (LPS, HSMM, Germline and MTAB). The results show the feasibility and high predictive accuracy of TiC2D in determining cellular states and inferring cellular trajectories. By comparing TiC2D to the state-of-the-art methods, we found that TiC2D can improve the accuracy of inferred trajectories from single cell transcriptome. The good performance of TiC2D implies that the consensus clustering strategy and the average similarity of cluster are effective in improving the pseudo-time trajectory inference. Furthermore, the reconstructed trajectories enable us to identify key genes involved in cell fate determination and to obtain new insights about their roles in dynamic cellular programs.

Funding

This work was supported in part by the National Natural Science Foundation of China (61772128, 61972100, 61772367), the Fundamental Research Funds for the Central Universities (2232016A3-05) and Shanghai Natural Science Foundation (18ZR1414400, 19ZR1402000).
Conflict of Interest: none declared.

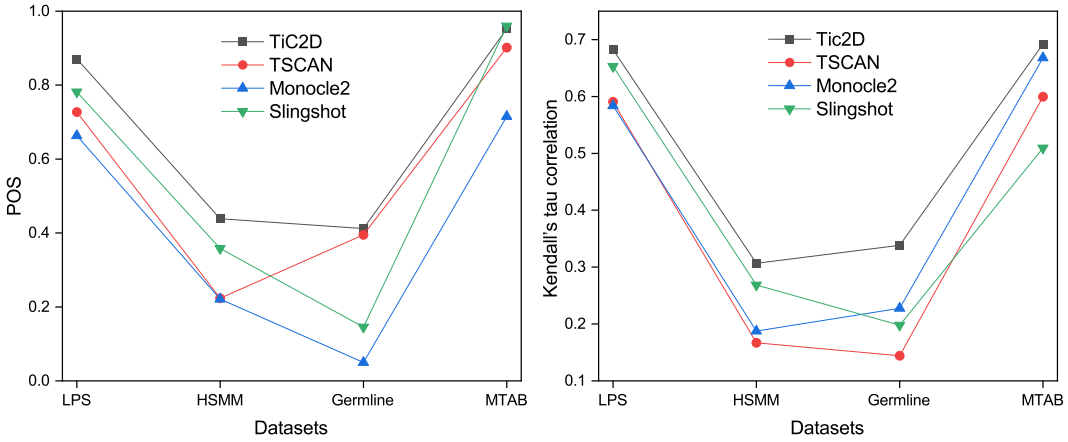


Fig. 5. Comparison of TiC2D with state-of-the-art algorithms for pseudo-temporal trajectory inference. The accuracy is measured by Pseudo-temporal Ordering Score (POS).

REFERENCES

[1] G. Balázsi, A. van Oudenaarden, and J. J. Collins, "Cellular decision making and biological noise: from microbes to mammals," *Cell*, vol. 144, no. 6, pp. 910–925, 2011.

[2] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan *et al.*, "Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, p. 1131, 2013.

[3] S. Semrau, J. E. Goldmann, M. Soumillon, T. S. Mikkelsen, R. Jaenisch, and A. Van Oudenaarden, "Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells," *Nature Communications*, vol. 8, no. 1, p. 1096, 2017.

[4] T. Baslan and J. Hicks, "Unravelling biology and shifting paradigms in cancer with single-cell sequencing," *Nature Reviews Cancer*, vol. 17, no. 9, p. 557, 2017.

[5] M. Guo, Y. Du, J. J. Gokey, S. Ray, S. M. Bell, M. Adam, P. Sudha, A. K. Perl, H. Deshmukh, S. S. Potter *et al.*, "Single cell rna analysis identifies cellular heterogeneity and adaptive responses of the lung at birth," *Nature Communications*, vol. 10, no. 1, p. 37, 2019.

[6] M. B. Woodworth, K. M. Girskis, and C. A. Walsh, "Building a lineage from single cells: genetic techniques for cell lineage tracking," *Nature Reviews Genetics*, vol. 18, no. 4, p. 230, 2017.

[7] F. Costa, D. Grün, and R. Backofen, "Graphddp: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge," *Nature Communications*, vol. 9, no. 1, p. 3685, 2018.

[8] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature Biotechnology*, vol. 32, no. 4, p. 381, 2014.

[9] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[10] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, vol. 34, no. 6, p. 637, 2016.

[11] L. Haghverdi, M. Buettner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature Methods*, vol. 13, no. 10, p. 845, 2016.

[12] H. Chen, L. Albergante, J. Y. Hsu, C. A. Lareau, G. L. Bosco, J. Guan, S. Zhou, A. N. Gorban, D. E. Bauer, M. J. Aryee *et al.*, "Single-cell trajectories reconstruction, exploration and mapping of omics data with stream," *Nature Communications*, vol. 10, no. 1, p. 1903, 2019.

[13] Z. Ji and H. Ji, "Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis," *Nucleic Acids Research*, vol. 44, no. 13, pp. e117–e117, 2016.

[14] J. Chen, A. Schlitzer, S. Chakarov, F. Ginhoux, and M. Poidinger, "Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development," *Nature Communications*, vol. 7, p. 11988, 2016.

[15] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature Methods*, vol. 14, no. 10, p. 979, 2017.

[16] M. Guo, L. E. Bao, M. Wagner, A. J. Whitsett, and Y. Xu, "Slice: determining cell differentiation and lineage based on single cell entropy," *Nucleic Acids Research*, vol. 45, no. 7, pp. e54–e54, 2017.

[17] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome Biology*, vol. 20, no. 1, p. 59, 2019.

[18] Z. Chen, S. An, X. Bai, F. Gong, L. Ma, and L. Wan, "Densitypath: an algorithm to visualize and recon-

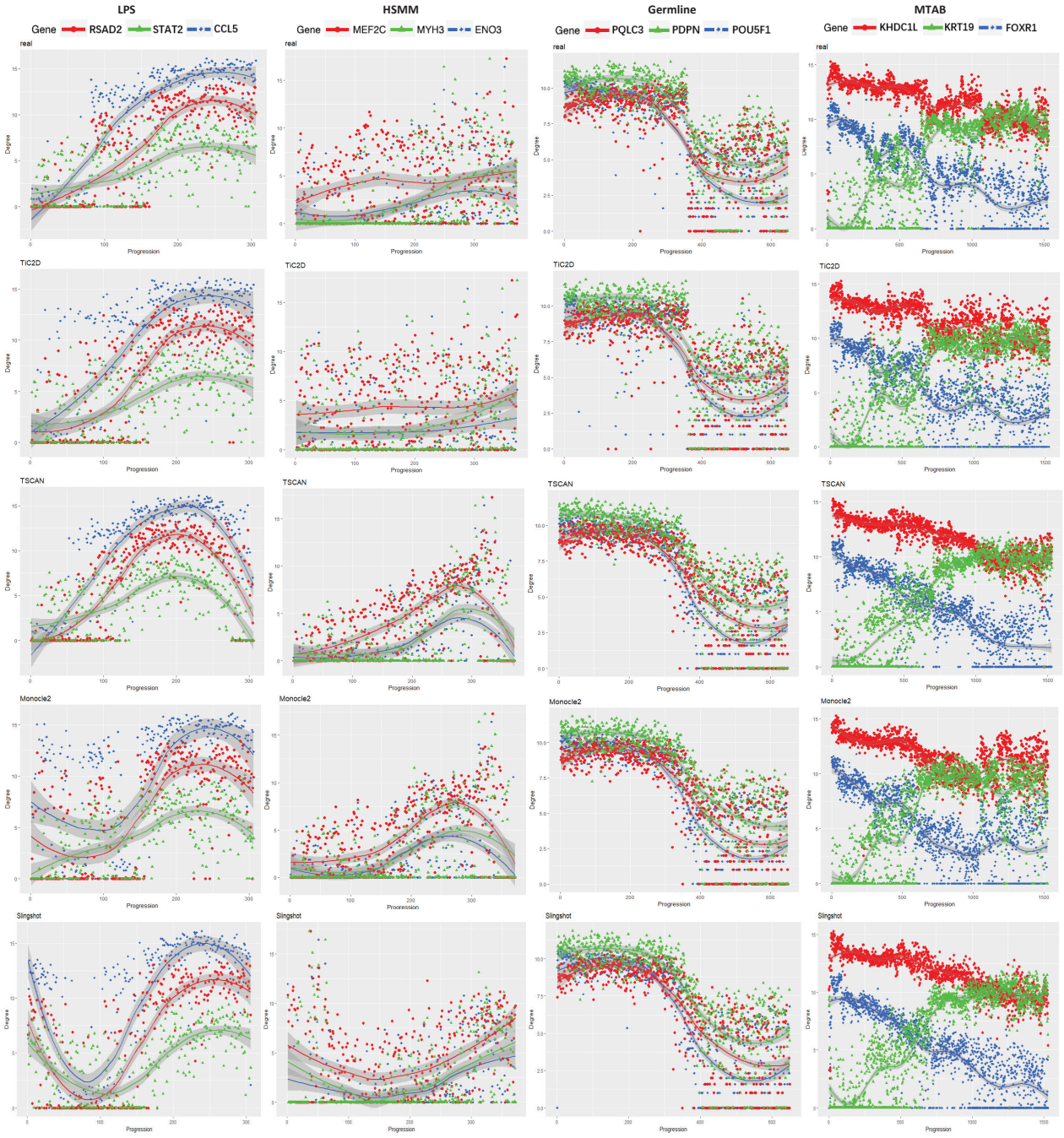


Fig. 6. The gene expression profiles of the marker genes from each dataset along the pseudo-time. The horizontal axis represents the pseudo-time, and the vertical axis represents the expression levels of the marker genes of each cell. The solid red line denotes a LOESS fit.

struct cell state-transition path on density landscape for single-cell rna sequencing data," *Bioinformatics*, 2018.

[19] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[20] I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," *Journal of Educational Measurement*, vol. 40, no. 3, pp. 277–280, 2003.

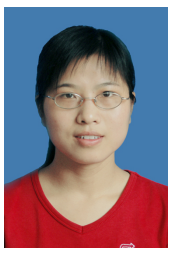
[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, p. 133, 2015.

[23] H. Liu, R. Zhao, H. Fang, F. Cheng, Y. Fu, and Y.-Y. Liu, "Entropy-based consensus clustering for patient stratification," *Bioinformatics*, vol. 33, no. 17, pp. 2691–2698, 2017.

[24] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green *et al.*, "Sc3: consensus clustering of single-cell rna-seq data," *Nature Methods*, vol. 14, no. 5, p. 483, 2017.

[25] I. Amit, M. Garber, N. Chevrier, A. P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J. K. Grenier, W. Li, O. Zuk *et al.*, "Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses," *Science*, vol. 326, no. 5950, pp. 257–263, 2009.
[26] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature Biotechnology*, vol. 37, no. 5, p. 547, 2019.
[27] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S. P. Panula, S. Codeluppi, A. P. Reyes, S. Linnarsson, R. Sandberg, and F. Lanner, "Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos," *Cell*, vol. 165, no. 4, pp. 1012–1026, 2016.
[28] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology*, vol. 33, no. 5, pp. 495–502, 2015.
[29] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
[30] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[31] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, no. 1, p. 477, 2018.
[32] R. Cannoodt, W. Saelens, D. Sichien, S. Tavernier, S. Janssens, M. Guilliams, B. N. Lambrecht, K. De Preter, and Y. Saeys, "Scorpius improves trajectory inference and identifies novel modules in dendritic cell development," *bioRxiv*, p. 079509, 2016.
[33] S. Watanabe, "A widely applicable bayesian information criterion," *Journal of Machine Learning Research*, vol. 14, no. Mar, pp. 867–897, 2013.
[34] J. A. Pulit-Penaloza, S. V. Scherbik, and M. A. Brinton, "Activation of oas1a gene expression by type i ifn requires both stat1 and stat2 while only stat2 is required for oas1b activation," *Virology*, vol. 425, no. 2, pp. 71–81, 2012.



Yanglan Gan is an associate professor in the school of computer science and technology at Donghua University, Shanghai, China. She received the Ph.D. degree in computer science from Tongji University in 2012, China. She has worked as a Visiting Scholar in the Department of Computer Science and Engineering at Washington University in St. Louis from 2009 to 2011, USA. Her research interests include bioinformatics, data mining, and Web services. She has published more than 40 papers on international journals and conferences, including Bioinformatics, IEEE/ACM TCBB, BMC Bioinformatics, BMC Genomics, Knowledge-based Systems, and Soft Computing. She served as a program committee member on BIBM 2019 and GIW 2018. She worked as a reviewer for a variety of international journals and conferences, such as BMC Bioinformatics, Journal of Molecular Biology, Knowledge-based Systems.

Ning Li is currently a master student in the School of Computer Science and technology, Donghua University, China. Before that, he received a Bachelor degree in Computer Science and Technology at Shanghai Polytechnic University, 2017. His research interests include bioinformatics and machine learning.

Cheng Guo is currently a master student in the School of Computer Science and technology, Donghua University, China. Before that, he received a Bachelor degree in Taiyuan University of Technology, 2020. His research interests include data mining and bioinformatics.

Guobing Zou is an Associate Professor and Dean of the Department of Computer Science and Technology, Shanghai University, China. He received his PhD in Computer Science from Tongji University, Shanghai, China, 2012. His current research interests focus on data mining, intelligent algorithms and services computing. He has published around 70 papers on premier international journals and conferences,including Information Sciences, Expert Systems with Applications, Knowledge-Based Systems, IEEE Transactions on Services Computing, AAAI, ICWS and ICSOC.

Jihong Guan received the bachelor's degree from Huazhong Normal University in 1991, the master's degree from the Wuhan Technical University of Surveying and Mapping (merged into Wuhan University in 2000) in 1998, and the Ph.D. degree from Wuhan University in 2002. She is currently a Professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. Her research interests include databases, data mining, distributed computing, bioinformatics, and geographic information systems. She has extensively published more than 300 papers in domestic and international Journals (including Nature Communications, IEEE TKDE/ TITS/TSC/TGRS/TCBB, NAR, Bioinformatics) and conferences (including AAAI, ICDE, VLDB, SIGIR, RECOMB, DASFAA).

Shuigeng Zhou received the Bachelor's degree from the Huazhong University of Science and Technology, in1988, the Master's degree from the University of Electronic Science and Technology of China, in 1991, and the PhD degree in Computer Science from Fudan University, Shanghai, China, in 2000. He is a full professor in the School of Computer Science, Fudan University. His research interests include data management, data mining, machine learning, and bioinformatics. He has extensively published in domestic and international journals and conferences.