Single Cell Trajectory Inference Based on Mouse Embryo Dataset

Xiaoxuan Chen *

Department of life science, Jilin University, Changchun, China, 130012

* Corresponding Author Email: chenxx1321@mails.jlu.edu.cn

Abstract. Constructing cell differentiation trajectories helps understand the development process of normal tissues or provides pathology-related information. This work analyses data from a mouse model of fetal growth restriction. In data processing, this work encodes for genes, uses PCA dimension reduction to reduce the cost of calculating, and uses the Louvain algorithm to cluster cells into seven different types including DN, DP-L, DP-M1, DP-M2, DP-M3, T-cell, and unknown, and then infer cell differentiation trajectories by DDRTree. By analyzing cell differentiation trajectories, the cell differentiation process can be known. Besides that, it can be found that the growth of DP cells can be divided into different steps due to differences in gene expression level.

Keywords: Trajectory Inference, Single Cell Inference Dataset, Dimension Reduction, Clustering, T-cell.

1. Introduction

The data we used for single-cell trajectory comes from single-cell sequencing. In the past decade, bulk RNA sequencing technology has been widely used to study gene expression patterns at the population level. This technique performs high-throughput sequencing of large amounts of RNA isolated from cells. In recent years, with the continuous innovation of sequencing technology, various protocols based on single-cell RNA sequencing (scRNA-seq) have been proposed, which promotes the study of cell type and specific cell type marker genes and cell heterogeneity at single-cell resolution. Currently, most mainstream high-throughput RNA sequencing platforms are based on microfluid-based or droplet-based technologies.

After the single-cell RNA sequencing is completed, the quality control (QC) of the original sequencing data needs to be carried out, and then the data are aligned to the reference genome, and the expression matrix is obtained by expression quantification of the data. The rows and columns of the expression matrix represent genes (transcripts) and cells, respectively, and each element of the matrix represents the expression level of a particular gene in the corresponding cell. Typical analysis of single-cell datasets includes processes such as cell typing clustering, differential gene analysis between different cell populations, trajectory inference of cell lineages (or pseudo time series analysis), and regulatory network reconstruction [1].

Constructing cell differentiation trajectories from single-cell transcriptomics data or proteomics data by single-cell trajectory inference methods helps understand the development process of normal tissues or provides pathology-related information.

There are more than 70 methods for single-cell trajectory inference. To deal with the high-dimensional noise in scRNA-seq data, feature dimension reduction and feature selection algorithms are widely used in single-cell trajectory inference. Monocle and Monocle2 algorithms have been used to infer various types of cell differentiation trajectories, including the inference of some neuron lineage trajectories, the division of progenitor cells and stem cells in the hematopoietic system, and the development of fetal discs. Then SCUBA (single-cell clustering using bifurcation analysis) uses t-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensionality of the data, and then fits a smooth curve to the reduced data. SCUBA has been used to calculate key genes in the differentiation of stem cells in the mouse small intestine. Slingshot takes the dimensionality reduction data and the clustering results as input to construct the MST, and then fits smooth curves on each branch of the MST to further refine the tree. Finally, all cells are projected onto their closest curve,

resulting in an ordered lineage trajectory with branches. Slingshot has recently been used to predict cell fates and branch points in olfactory stem cell lineage trajectories [2]. Figure 1 shows 7 types of possible topology for a trajectory. The Connected graph is tantamount to that there is a circle inside the Tree, and the Disconnected graph is tantamount to that there is a break inside connection.

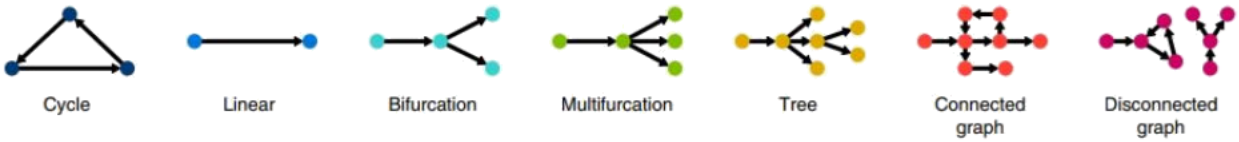


Figure 1. The 7 types of possible topology for a trajectory

2. Data description

The recourse of sequencing data comes from a mouse model of fetal growth restriction [3]. The dataset includes 15395 rows, meaning the number of genes, and 8605 columns, representing the number of cells. For the single-cell sequencing dataset, we searched for results from three documents' integration. The three documents include data about cells, genes, and matching.

3. Data processing

3.1. Gene coding

For convenience to research and comprehend, we turn complex gene ID, including numbers and letters, into gene symbols.

3.2. PCA dimension reduction

PCA (Principal Component Analysis) stands as a widely applied algorithm for reducing data dimensionality. The fundamental idea of PCA is to map the n-dimensional features to k-dimensions. These new orthogonal features (k-dimensions), termed principal components, are orthogonal features crafted through the reconstruction of the original n-dimensional features. PCA's core function is to sequentially find a set of orthogonal axes closely related to the data itself from the original space. The initial axis is chosen in the direction of maximum variance present in the original data, followed by the selection of subsequent axes in orthogonal planes, each chosen to maximize variance. This process continues until a total of n such axes are derived.

We use this method to reduce dimensions from 8605 to 210. It is a good method to reduce calculation costs by reducing the items that would be analyzed. Meanwhile, because of the sparsity of cells' data, there are many "0" in the matrix of genetic expression. PCA can reduce the "0" , making it convenient for the following analysis.

As depicted in Figure 2, we found that the variance explanation rate approaches 0 when the dimension is over 210. That means these data have little difference and can be abandoned. The rationale behind this lies in the acquisition of new axes through the aforementioned method is that the majority of variance is included within the initial k axes, while the subsequent axes contribute almost negligible variance. Consequently, disregarding the surplus axes and retaining solely the first k axes, which harbor the majority of the variance, is akin to selectively preserving dimensional features crucial for variance while discarding those with nearly zero variance. This approach effectively achieves the reduction of data feature dimensionality.

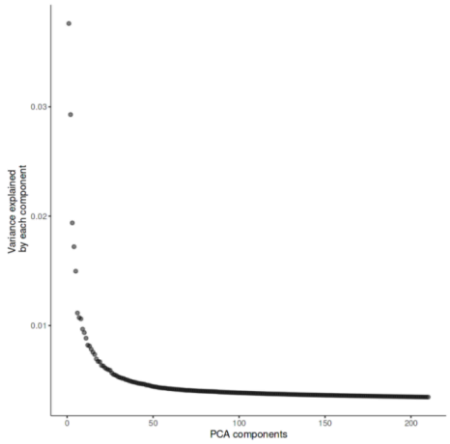


Figure 2. Result of PCA

3.3. Clustering

This part aims to divide cells into different types. Then, we can research how those different types of cells transform into each other. The method we used for the cluster is Louvain based on a technology called “community detection”. Community detection is frequently employed to comprehend the intricate structure of large networks. One of the most popular algorithms for uncovering community structure is the so-called Louvain algorithm [4].

Louvain algorithm includes several steps. First, regard every node in the community, which means an organization with potential relationships, as a new community, and combine it with its neighbor community, which makes the whole community own biggest increment of modularity, which could measure the quality of nodes’ grouping. As a result, the whole dataset is divided into different community preliminary. Second, regard these communities as a new node. Repeat step 1, until getting a satisfactory result.

The result of clustering is shown in Figure 3. Each point means a cell from a mouse embryo. They have been clustered into different communities. We can see that the number of DP-M2 is the most, followed by DP-M1. Lower than them, the number of T-mat and DN is similar, followed by DP-L. Except for cells unknown, the number of DP-M3 is the lowest. The number of cells is likely related to the importance of the growth of mouse embryos.

Different color means different types of cells. In Figure 3, DP-M1 stands for double positive-middle T-cell (1st cluster), DP-M2 stands for double positive-middle T-cell (2nd cluster), 'T-mat' stands for mature T-cell, DN stands for double negative-early T-cell, DP-L stands for double positive-late middle T-cell, DP-M3 stands for double positive-middle T-cell (3rd cluster). DN lacking CD4 and CD8 receptors and DP with CD4 and CD8 receptors are two different T-cell growth stages. “Unknown” stands for the cells that haven’t been found or unidentified.

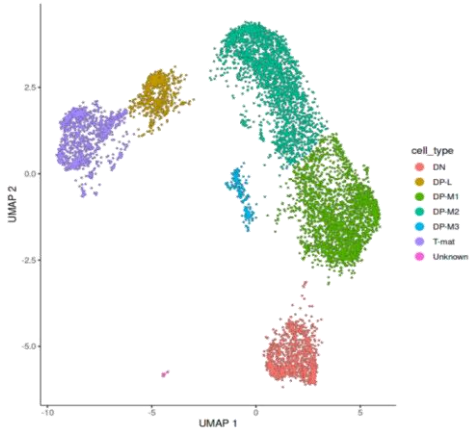


Figure 3. Result of clustering

3.4. Confirming and screening of marker genes

Figure 4 shows that marker genes express differently in different clusters. A marker gene is a kind of gene whose sequence or function has been known clearly and is only expressed at a high level among specific types of cells. We can identify cell types of each cluster by checking marker genes' expression.

In figure 4, the area of the circles stands for the number of genes, while the color of the circles represents the amount of gene expression. We can see that the amount of Pcdhgc4 and Pcdhga5 occupies most in cluster 7, while the amount of Gm5559 and GM10359 occupies most in cluster 6. Ccr9 and Cd8b1 express most in cluster 5, while Cd52 and Tmsb10 express most in cluster 3. The amount of Plac8 and Il2ra is the most in cluster 4. Mki67 and Hmgb2 express most in cluster 2, while Hmgb2 express higher obviously. In cluster 1, expression and amount of Pclaf and Npm1 are slightly higher than others, but not very specific.

Overall, genes are not as specific as expected, appearing across the whole sample. We should plot all those marker genes and check how they appear in the bigger picture. Compared with the illustrate situation of gene expression in an abstract way in Figure 4, Figure 5 shows what cell types correspond to the genes expressed in each cluster specifically, even specific to each cell.

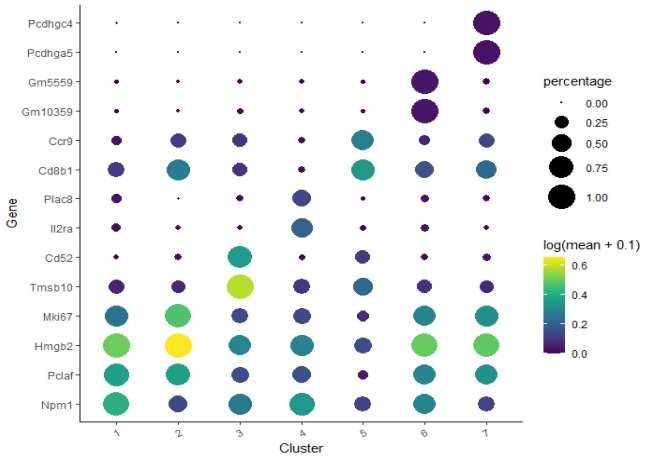


Figure 4. A plot of the expression and fraction of cells that express found markers in each group

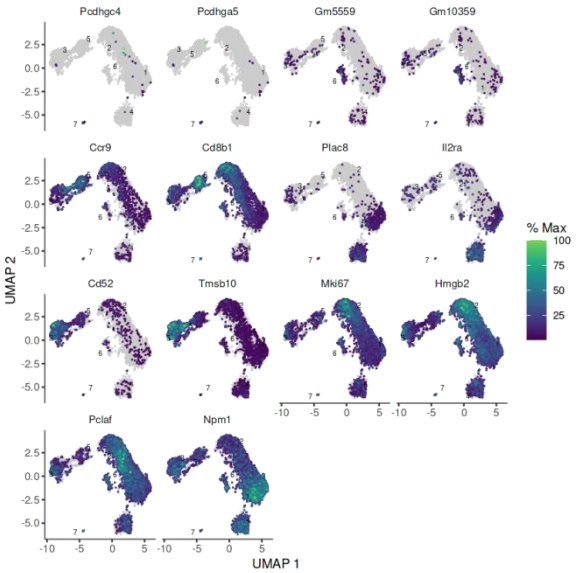


Figure 5. Plots of the expression of marker genes across the sample

4. Cell differentiation inferring trajectories

After clustering, we've gotten different cell types. In this part, we will research how different types of cells transform into each other as pseudo time goes by.

In this section, we leverage the reversed graph embedding (RGE) technique to acquire insights into the graph structure characterizing a single-cell experiment. RGE, being a machine-learning approach, simultaneously grasps a principal graph illustrating the cell trajectory. Additionally, it learns a function capable of mapping points on the trajectory, embedded in lower dimensions, back to the original high-dimensional space. [5].

There is a set of latent points $Z = \{z_1,..., z_N\}$ and an undirected graph G. N stands for the number of cell numbers, and G builds connection among these latent points. The graph $G = (V,\varepsilon)$ includes several vertexes $V = \{V_1,..., V_N\}$ and a set of weighted, undirected edges ε, where each V_i corresponds to latent point z_i. We learn them by defining f_G as the projection from z_i to some point in the high-dimensional space, and b_{ij} denotes the weight of the edge (V_i, V_j), which represents the connectivity between z_i and z_j.

$$\min_{G\in G_b}\min_{f_G\in F}\min_{Z}\sum_{(V_i,V_j)\in\varepsilon} b_{i,j}\,\|f_G(z_i)-f_G(z_j)\|^2 \tag{1}$$

G_b is a set of feasible graph structures with parameters $\{b_{i,j}, \forall i, j\}$, and F is a set of functions mapping a latent, low-dimensional point to a point in the original, high-dimensional space. Graph structures learned by this method in the latent space. However, this method does not measure the deviations of the latent points to the observed data. So, we need to optimize it.

$$\min_{G\in G_b}\min_{f_G\in F}\min_{Z}\sum_{i=1}^{N}\|x_i-f_G(z_i)\|^2+\frac{\lambda}{2}\sum_{(V_i,V_j)\in\varepsilon} b_{i,j}\,\|f_G(z_i)-f_G(z_j)\|^2 \tag{2}$$

λ is a parameter that adjusts the relative strength of these two summations.

In practice, some constraints on Gb and fG are important to realize reversed graph embedding. There have developed several methods aimed at implementing RGE, including the simple principal tree (SimplePPT), principal L1 graph, dimensionality reduction via learning a tree (DRTree), and discriminative dimensionality reduction via learning a tree (DDRTree) algorithms [6].

SimplePPT represents the pioneering RGE technique designed to acquire a tree structure for delineating a collection of observed data points. Not only the tree can be learned in the original space, but also the space after dimensionality reduction as well. fG(zi) is optimized from two separate sets of variables to one single variable. Meanwhile, the empirical quantization error serving as the measurement between the fG(zi) and its corresponding observed points xi replaces the loss function in the RGE.

The L1 graph is characterized by its sparsity, operating on the premise that each data point (or cell) is associated with a limited number of neighborhoods. Among these neighborhoods, the minimum number of points that span a low-dimensional affine subspace passes through that point. Distinguishing itself from SimplePPT, this approach expresses the optimization task as a linear programming problem in order to acquire knowledge of the graph.

In contrast, DDRTree excels at dimensionality reduction while concurrently learning the trajectory, without presuming that the graph is confined to the input space. Additionally, it dispenses with the requirement for only one node in the graph per data point, resulting in significant algorithmic acceleration and a reduced memory footprint.

This approach employs DDRTree, which yields a primary tree depicting the centroids of cell clusters in low dimensions. The aim is to understand the primary tree that characterizes a single-cell experiment and subsequently map each cell (represented by latent points Z) onto the closest location on the tree (constituted by principal points Y forming the principal graph). By selecting a tip of the tree as the root, we traverse the tree from the root, compute the geodesic distance from each cell to the root cell, and use this distance as its pseudotime. This is done concurrently with assigning the cell

to a specific branch or segment. For latent points not in proximity to the 'tip' principal points, the method identifies the nearest line segment on the principal tree and projects the point onto the nearest location on that segment. We can define a vector between a cell c = (c1, c2, …), where c1, c2, … stand for the coordinates of the cell in the latent space, and the nearest principal point A by $\overrightarrow{\mathrm{Ac}}$, while the line segment between the nearest two principal points (A = (A1, A2,…) and B = (B1, B2,…)) is $\overrightarrow{\mathrm{AB}}$. We can calculate t as

$$t = \frac{\overrightarrow{Ac} \cdot \overrightarrow{\mathrm{AB}}}{\| AB \|} \tag{3}$$

The projection can be calculated as:

$$p = \begin{cases} A & if \quad t < 0 \\ B & if \quad t > 1 \\ A + t \cdot \overrightarrow{AB} & if \; 0 \le t \le 1 \end{cases} \tag{4}$$

$A + t \cdot \overrightarrow{AB}$ is a projection point. We obtain it by orthogonally projecting the latent point nearing the tip principal points, which is in proximity to the tip principal points, onto the line segment extending from the tip principal point to its nearest neighbor principal point in the graph. After that, we measure the distances between all projection points and construct a minimal spanning tree (MST). This MST serves as a basis for assigning pseudotime to each cell based on the projection points[5].

Figure 6 illustrates the direction of cell differentiation in pseudo-time. The color stands for pseudo-time nodes. The differentiation begins with the dark blue area below and a yellow area at the end. Figure 7 shows the relationship of differentiation among specific cell types.

We can see that the trajectory is linear overall, and DN experience steps of DP-M1, DP-M3, DP-M2, DP-L, and finally differentiate to mature T-cell. Before the step of DP-L, we can see many branches during the differentiation of DP cell. That speculated to be the cells that killed in positive selection, which induces T-cell that are unable to combine with major histocompatibility complex(MHC) apoptosis, and negative selection, which induces T-cell that combine with MHC or autoantigen too strong apoptosis. These process guarantee mature T-cell identify MHC submitted by autologous but not by others', and get tolerance to autoantigen [7]. There are a few cells tagged as Unknown and can't be connected with others. They speculated to be noise or error, which can be ignored.

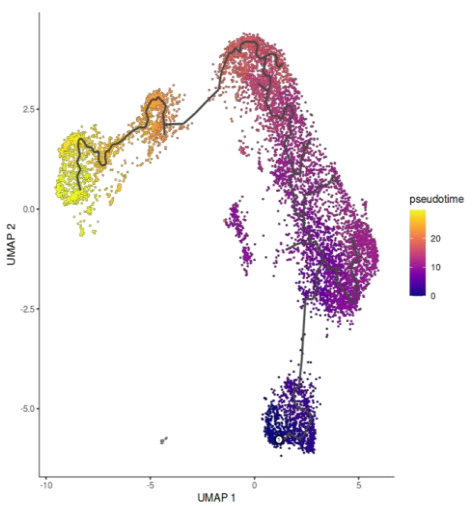


Figure 6. Cells ordered in pseudotime

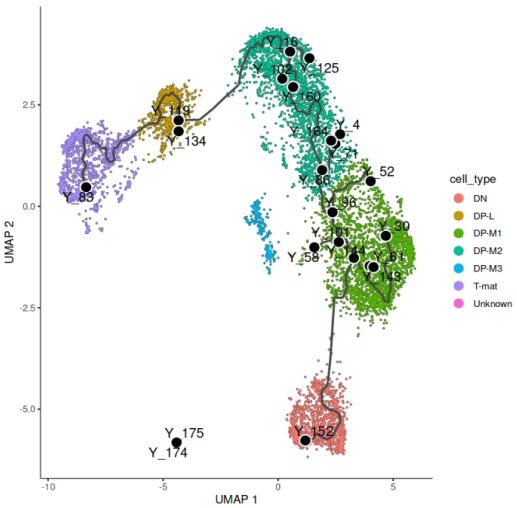


Figure 7. Learned trajectory graph

5. Gene expression differential analysis

In figure 8, the black points stand for the median, and the area around the stand for the distribution of data. We can see that some genes express differently in different cell types, while others like Hba-a1 act similarly among these cell types. Obviously, Aif1 and Il2ra express the highest in DN. Cd4, Cd8a and Cd8b1 express highest in DP-L. Itm2a expresses highest in T-mat.

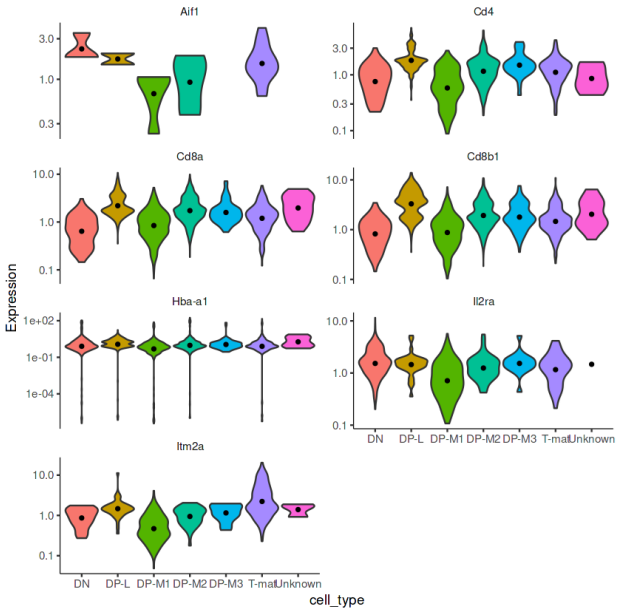


Figure 8. Violin plots of the expression of the specified genes in each assigned cell type

The figure 9 shows how specified gene expression situations change in the pseudotime. Different colors mean different cell types, while abscissa means pseudotiom goes by from left to right. The genes shown in the figure all own different function.

Aif1 expresses allograft inflammatory factor 1 (AIF-1), a combination of actin protein and calcium, induced by cytokines and interferon, which can promote macrophage activation and the growth of and T cells [8]. We can see that Aif1 is expressed similarly among all these types of cells.

Cd4, expression of which witnesses a slightly increase when differentiated to the end of DP-M2, and a decrease when it comes to DP-L, express Cd4 molecule, which extracellular domain can combine with MHC-IIand TCR to mediates specific antigen presentation while cytoplasmic domain interaction with tyrosine kinase to mediates antigen signaling, and finally activate CD4+ T cell, trigger a series of immune responses [9].

Cd8a increases at the beginning of DP-M2, decreases in DP-L, and finally remains unchanged at about 27. It acts upstream of or within several processes, including cytotoxic T-cell differentiation and defense response to viruses.Cd8b1 increases at begin of DP-M2 and decreases from step of DP-L to T-mat. It is predicted to enable MHC class I protein binding activity and coreceptor activity, which regulate immune response and act upstream of or within the adaptive immune response.

Il2ra decreases at the beginning of DN and then remains unchanged. It expresses interleukin-2 receptor and enables interleukin-2 binding activity activation-induced regulation of T cell proliferation. IL-2 is a four-helix bundle cytokine composed of glycoproteins, which is mainly produced by the interaction of CD4+ and CD8+ T lymphocytes. IL-2 must be involved in the CD8+ memory T cell response in persistent T cell immunity [10]. Canonical interleukin-2 (IL-2) signaling via the high-affinity CD25-containing IL-2 receptor-Janus kinase (JAK) 1,3-signal transducer and activator of transcription 5 (STAT5) pathway is crucial for the development and maintenance of CD4+ CD25Hi Foxp3+ regulatory T cells (Tregs) that support immune homeostasis [11]. Hba-a1 expressed similarly on a high level relatively among all these types of cells. Meanwhile, Itm2a's expression remains unchanged before about 23 in DP-L and then increases gradually.

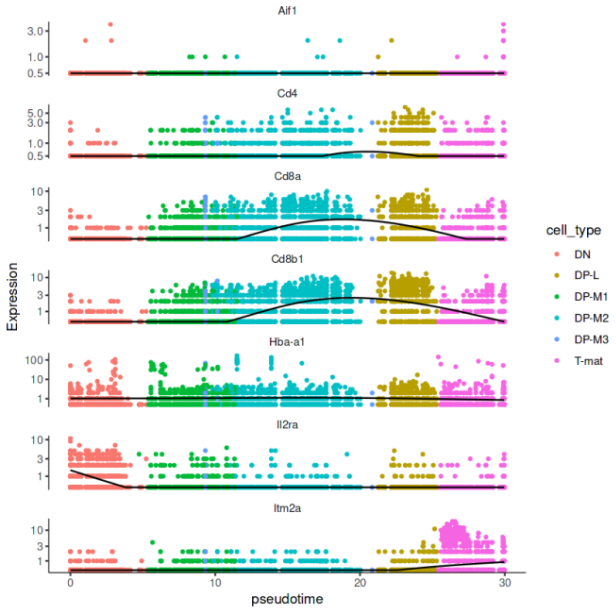


Figure 9. A visualisation of how specified genes change in the pseudotime

6. Conclusions

This research first reduces data dimension, clusters the cells into 6 different types, infers trajectories, and finally analyses gene expression differentiation. We divide DP into several different clusters due to its situation of differentiation, analyze the trajectory of differentiation from DN cell to T cell, and associate the function of genes with different cell types. It may provide a clearer prospect of T cell differentiation in mouse embryos.

Growth of T cell experiences several steps, including DN cell and DP cell. Meanwhile, genes express differently during the growth of DP cell. Expression products of these genes all play important roles in mature T-cell, and ensure T-cell differentiates to Tc, Th, etc, successfully when touch antigen. Aif1 promotes the growth of T-cell on similar degree during the whole pseudo time. CD4 and CD8 are both receptors on the cytomembrane of DP. After positive and negative selection, DP turns into a single positive cell (CD4-CD8+ or CD4+CD8-). They are all mature T cell. And the result, CD4 and CD8 express highly during the period of DP cells' growth, conforming to the aforementioned. Il2ra express highly in DN to produce more IL-2 receptor, and promotes the growth of DN.

References

[1] Shaoqiang Zhang, Lixuan Li, Linjuan Xie. Review of imputation methods for single cell RNA sequencing data [J]. Journal of Tianjin Normal University (Natural Science Edition),2023,43(2):1-10,73.

[2] Hongjian He, Yiting Yin, Jiang XIE. Single cell differentiation trajectory inference algorithm based on iterative feature selection [J]. Computer Science and Exploration,2023,17(7):1609-1621.

[3] Bacon WA, Hamilton RS, Yu Z, Kieckbusch J, Hawkes D, Krzak AM, et al. Single-cell analysis identifies thymic maturation delay in growth-restricted neonatal mice. Frontiers in immunology. 2018; 9:2523-

[4] Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019).

[5] Qiu Xiaojie, Mao Qi, Tang Ying, et al. Reversed graph embedding resolves complex single-cell trajectories[J]. Nature methods,2017,14(10):979-982

[6] Jinmiao Chen, Laurent Rénia, Florent Ginhoux.Constructing cell lineages from single-cell transcriptomes[J]. Molecular Aspects of Medicine,2018,59:95-113

[7] Guocheng Zhong, Xiaohong Li, Bo Zhu. Research progress of microRNA and T cell development [J]. Chongqing Med, 2012, 41(6): 612-614.

[8] Hufang Wang, Yuwei Yan, Xiuyan Wang, et al. Cervical cancer patients' serum TAGLN2 and AIF1 expression and the clinical diagnostic value [J]. Hebei medicine, 2023, (8): 1190-1193

[9] Jian Cui, He Xiao, Jiannan Feng, et al. Advances in functional epitopes and antagonists of CD4 [J]. Journal of Chinese Academy of Military Medical Sciences,2007,31(2):170-174.

[10] Jiaxin Liang, Zhongqin Luo, Dongsong Nie, et al. Research progress of Bai Jiesu2 [J]. Journal of Hunan Institute of Science and Technology (Natural Science Edition), 222,35(2):18-21.

[11] Sun Hao, Lee Ho-Sup, Kim Sarah Hyun-Ji, et al.IL-2 can signal via chemokine receptors to promote regulatory T cells' suppressive function[J]. Cell Reports, 2023 42 (8)