ALLSTATE: Hierarchical Clustering for Single Cells based on Non-linear Transition Embedding

Yating Lin[1], Minshu Wang[2,3], Wenxian Yang[4], and Rongshan Yu[1,3,4]

[1] School of Informatics, Xiamen University, Xiamen 361005, China
[2] School of Medicine, Xiamen University, Xiamen 361102, China
[3] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China
[4] Aginome Scientific, Xiamen, 361005, China
rsyu@xmu.edu.cn

Abstract. Single-cell RNA sequencing (scRNA-seq) provides critical insights into cellular diversity, essential for understanding complex biological dynamics. Traditional scRNA-seq analysis employs either unsupervised clustering methods or supervised learning-based approaches to interpret cells or cell clusters, but both lacks the flexibility to adjust the clustering resolution needed to fully capture the complex spectrum of cell types and states. On the other hand, hierarchical clustering can explore cell distribution structures across various resolutions without predefined cluster counts, thus overcoming the limitations of both unsupervised and supervised methods. Nevertheless, the high-dimensional nature of scRNA-seq poses significant analytical challenges to hierarchical clustering. In addition, as scRNA-seq data usually exhibit rich nonlinear structures in the high-dimensional space, linear dimension reduction methods such as Principal Components Analysis (PCA) are usually notable to reveal these structures for effective cell type analysis. This study introduces ALLSTATE, a novel single-cell data processing pipeline that combines non-linear transition embedding and hierarchical clustering in a computationally efficient manner. Our experiments demonstrate that ALLSTATE achieves satisfactory clustering performance and allows us to explore the connections between cellular hierarchies and cell types at multiple levels of resolution. Additionally, ALLSTATE further enables capturing complex cellular differentiation paths, offering a nuanced view of cellular heterogeneity with performance comparable to mainstream methods.

Keywords: scRNA-seq, Clustering, Non-linear transition embedding, Cellular differentiation path.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful technique to reveal the heterogeneity and diversity among cell populations, facilitating a deep comprehension of intricate biological phenomena [1–4]. A major analytical challenge for scRNA-seq data analysis is to associate each high-dimensional cell phenotype to a previously identified cell type when interpreting large quantities of single-cell data, which is also crucial for other downstream applications such as linking tumor microenvironment composition to patient outcomes [5]. Furthermore, single-cell studies usually require identifying the potential cell status evolution or cell development trajectory through trajectory inference analysis, which constructs cell evolution trajectory from the gene expression change patterns with the aid of complementary knowledge, such as RNA velocity, timepoint information, or other types of omics data [6, 7]. Typically, trajectory inference requires researchers to perform clustering on the original single-cell data first to identify stable cell states in the data as starting points and afterward connect these states to form a trajectory.

Cell type identification is usually done by cluster-then-annotate methods [8–11], whereby discrete cell types are identified through unsupervised clustering analysis, which are then annotated based on expressions of marker genes on different clusters. After cell type identification, researchers can manually select one or more clusters for further refinement based on prior knowledge. The accuracy of clustering is significantly affected by researchers' subjective choice of resolution. Moreover, the discrete selection of resolutions limits their capability to fully capture the complexities of cellular differentiation processes. Alternatively, supervised clustering methods leverage machine learning algorithms trained on pre-labeled data for cell type identification [12–15], thus enhancing pattern recognition and accuracy. Yet their dependency on training data poses limits on adaptability and generalization capabilities. In particular, similar to unsupervised clustering methods, their ability to discern different cell types is limited by the resolution at which datasets are annotated.

To identify the hierarchical structures of scRNA-seq data, hierarchical clustering stands out as a promising alternative due to its inherent ability to process data with layered structures [16]. These methods systematically build a dendrogram, mapping data points across a continuum of similarity, thereby unveiling detailed insights into their intertwined relationships and facilitating adaptive resolution adjustments tailored to distinct analytical goals. This strategy effectively eliminates the need for predetermined cluster counts, making it suitable for applications where the number of clusters is indeterminate or fluctuating. Moreover, it offers a flexible framework for data exploration across various levels of abstraction, adeptly navigating the intricacies of multi-layered structural analyses that traditional clustering techniques, both unsupervised and supervised, often struggle to handle.

Unfortunately, the direct application of hierarchical clustering to scRNA-seq data encounters a number of challenges. Firstly, the inherent high-dimensional nature of scRNA-seq data leads to a substantial computational burden. Secondly, the sparsity of data makes it difficult to find a representative metric that can accurately measure the differences between cells. Such a metric is crucial for the efficacy of hierarchical clustering. A potential solution to the above issues is the application of appropriate dimension reduction techniques to process scRNA-seq data, with the aim of obtaining an effective representation of high-dimensional data in a low-dimensional space. However, traditional linear dimension reduction methods like PCA [17], as well as model-embedded dimension reduction approaches [18, 19] that integrate dimension reduction within data processing models, often struggle to capture the rich characteristics of scRNA-seq data due to its complex, non-Euclidean manifold-like structure in high-dimensional space.

To address these issues, we propose ALLSTATE, a hier**A**rchica**L** c**L**u**ST**ering pipeline b**A**sed on non-linear **T**ransition **E**mbedding in manifold space. This pipeline combines the power of non-linear transition embedding to extract diverse information pertaining to cell differentiation paths, with the capability of hierarchical clustering methods to decipher the heterogeneous state of cells in multiple resolutions. The proposed method demonstrates the ability to delineate the global structure of single-cell data without pre-defined resolution. In addition, it also provides comparable performance to the current mainstream scRNA-seq clustering methods when compared at pre-defined resolution.

2 Methodology

An overview of ALLSTATE is shown in Fig. 1. To delineate the structures of single-cell data at multiple resolutions through a unified manner, ALLSTATE first performs diffusion condensation to identify the hidden affinity-preserving manifold embedding of high-dimensional single cell data using PHATE [20]. Cell subtypes of different granularity are then identified by using simple hierarchical clustering on the identified manifold embedding. For clusters of single cells that form continuous distributions on the projected manifold, ALLSTATE further identifies their expression evolution trajectories by using the PAGA [21] algorithm based on their structures identified at finer resolutions, from which potential continuous cell state evolution or cell development trajectory can be identified.

2.1 Non-linear Transition Embedding

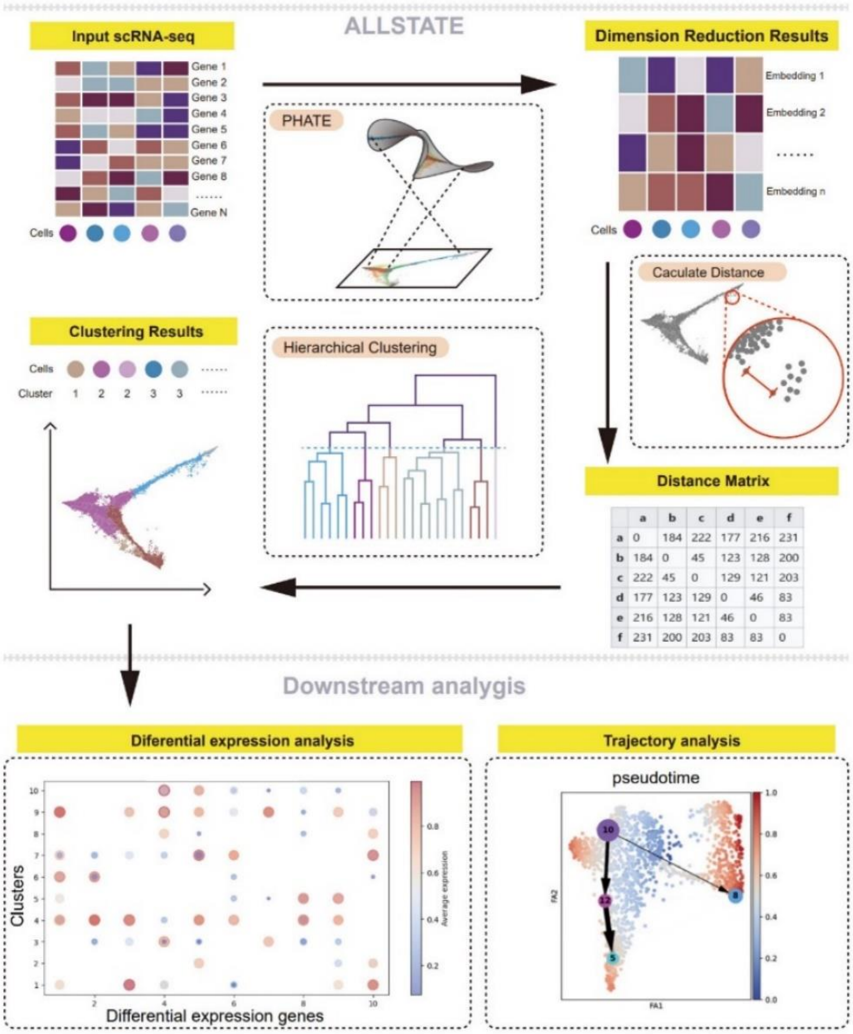
To capture the complex structure of scRNA-seq data, we employ the PHATE method. Briefly, starting with a high-dimensional expression matrix $X \in R_{m \times n}$, where m represents genes and n for cells, we construct a k-nearest neighbors (k-NN) graph. Cell-to-cell affinities are calculated using a Gaussian kernel:

$$\boldsymbol{A}_{ij} = exp\left(-\frac{d(x_i,x_j)^2}{\epsilon}\right) \tag{1}$$

where $\mathbf{A}_{ij}$ denotes the affinity between cells i and j, with $d\left(\mathbf{x}_i, \mathbf{x}_j\right)$ quantifying the Euclidean distance between their gene expression profiles. The parameter ϵ, determines the radius (or spread) of neighborhoods captured by this kernel. This forms an affinity matrix $\mathbf{A}$, which is used to generate a diffusion operator $\mathbf{D} = \mathbf{A} \cdot \mathbf{A}^T$. The operator captures the global structure of the input data through a diffusion process. By performing spectral decomposition on $\mathbf{D}$, we obtain eigenvectors $\mathbf{v}_i$ and corresponding eigenvalues λ_i:

$$\mathbf{D}\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{2}$$

Dimensionality reduction is achieved by selecting principal eigenvectors based on their eigenvalues, arranged in descending order. This selection, $\mathbf{X}_{\text{reduced}} = \mathbf{V}_{\text{selected}}$, provides a reduced representation that preserves the intrinsic geometry of data.



	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

Fig. 1. Overview of ALLSTATE.

2.2 Distance Calculation in Manifold Space

Following dimensionality reduction to obtain $\mathbf{X}_{\text{reduced}}$, we calculate the Euclidean distance between any two cells i and j in the reduced space as:

$$d_{\text{reduced}}(i,j) = \sqrt{\sum_{k=1}^{r}\left(x_{ik}^{\text{reduced}} - x_{jk}^{\text{reduced}}\right)^2} \tag{3}$$

where r represents the dimensions post-reduction, and x_{ik}^{reduced}, x_{jk}^{reduced} are the coordinates of cells i and j. This generates a distance matrix D_{reduced} essential for subsequent analyses.

2.3 Hierarchical Clustering

For structure exploration, we apply agglomerative hierarchical clustering to $D_{reduced}$. Initially treating each point as an individual cluster, we iteratively merge them based on the shortest distance, using linkage criteria such as single, complete, average linkage, or Ward's method, detailed as follows.

$$Single\ Linkage: \quad d_{\text{SingleLinkage}}\left(C_i, C_j\right) = min(d_{ij}): x \in C_i, y \in C_j \tag{4}$$

$$CompleteLinkage: d_{\text{CompleteLinkage}}\left(C_i, C_j\right) = max\left(d_{ij}\right): x \in C_i, y \in C_j \tag{5}$$

$$Average\ Linkage: \quad d_{\text{AverageLinkage}}\left(C_i, C_j\right) = \frac{1}{|C_i||C_j|}\sum_{x \in C_i}\sum_{y \in C_j} d_{ij} \tag{6}$$

$$Ward's\ Method: \quad d_{\text{Ward}}\left(C_i, C_j\right) = \frac{|C_i||C_j|}{|C_i|+|C_j|} \cdot d_{\mu_i \mu_j}^2 \tag{7}$$

Merging continues until reaching a single cluster or a predetermined condition. This process yields a dendrogram, providing insights into data organization and cluster identification.

2.4 PAGA

After dimensionality reduction and clustering, we apply PAGA for further trajectory analysis. PAGA constructs a k-nearest neighbors (k-NN) graph from reduced data, partitioning it into clusters based on hierarchical clustering outcomes. Connectivity between clusters is quantified with the PAGA connectivity score:

$$\phi\left(C_i, C_j\right) = \frac{O\left(C_i,C_j\right) - E\left(C_i,C_j\right)}{\sqrt{V\left(C_i,C_j\right)}} \tag{8}$$

where $O\left(C_i, C_j\right)$ represents the observed edge count, $E\left(C_i, C_j\right)$ represents the expected count under a random model, and $V\left(C_i, C_j\right)$ represents the variance. This measurement distinguishes meaningful connections between clusters.

PAGA then simplifies the graph, where nodes stand for clusters and edges indicate the strength of connectivity, illuminating cellular interactions and potential transitions between states.

3 Experiment Settings

3.1 Data Description

We conduct experiments on six scRNA-seq datasets. These scRNA-seq datasets consist of cells with known labels, and their details are given in Table 1. Note that the MCA_sub dataset is derived from the Mouse Cell Atlas 2.0, where it has been downsampled to 13,000 cells from the Adult subset. Subsequently, cells belonging to the cell_type2 category with fewer than 50 cells were removed.

Table 1. Details of the six scRNA-seq datasets.

Datasets	Cells	Genes	Level 1 Cell Types	Level 2 Cell Types
GSE144735	22414	33694	6	40
Zilionis_mouse_lung	6738	28205	6	18
Shekhar	26830	13166	5	18
Young	5685	33658	5	15
Stewart_Mature	22536	33694	26	32
MCA_sub	12707	39483	11	41

3.2 Baseline Methods

We compare ALLSTATE with the following seven baseline methods.

PCA+HC [28]: It's an application of the traditional hierarchical clustering method. It runs linear dimensionality reduction using PCA followed by hierarchical clustering in the low-dimensional space, and can obtain clustering results of continuous multiple resolutions.

PCA+Kmeans [29]: It combines the linear dimensionality reduction of PCA and the clustering capabilities of Kmeans, providing a solution that balances dimensionality reduction and clustering efficiency.

Leiden and Louvain [30]: These two methods are community clustering algorithms that focus on discovering modular structures in the data to optimize community delineation of clusters.

HGC [31]: A graph-based hierarchical clustering method which emphasizes the hierarchy and structure of data, aiming to improve the interpretability and biological significance of clustering results.

scDeepCluster [18]: A deep learning clustering method which combines the Zero-Inflated Negative Binomial (ZINB) model-based autoencoder with clustering loss, optimizing clustering while performing dimension reduction.

DESC [19]: An unsupervised deep learning algorithm that iteratively learns cluster-specific gene expression representation and cluster assignments for scRNA-seq analysis.

4 Results and Analysis

In this section, we first explore the influence of using different linkage criteria on model performance and select an optimal distance metric for subsequent experiments. Next, we compare ALLSTATE with seven current widely-used scRNAseq data clustering methods. We demonstrate the effectiveness of our method in improving the quality of data dimension reduction and its applicability in cell subpopulation identification through visualization. Finally, we apply the entire pipeline to re-explore the subtype branches of a specific cell type and perform trajectory inference to verify the evolutionary pathways between subtypes of a specific cell type.

4.1 Parameter Analysis

Table 2. Experimental results on different distance measures in hierarchical clustering.

Datasets	Linkage Criteria	NMI	ARI	V-Measure
GSE144735	Single	0.7300	0.3735	0.7008
	Complete	0.7231	0.4341	0.7227
	Average	**0.7704**	**0.5485**	**0.7633**
	Ward	0.7252	0.5256	0.7247
Zilionis_mouse_lung	Single	0.8416	0.9140	0.8270
	Complete	0.8314	0.8883	0.8264
	Average	**0.8733**	**0.9624**	**0.8769**
	Ward	0.7152	0.4156	0.7147
Shekhar	Single	0.4420	0.1325	0.3310
	Complete	0.7202	0.5619	0.7188
	Average	0.7630	**0.6022**	0.7701
	Ward	**0.7712**	0.5934	**0.7751**
Young	Single	0.4864	0.3745	0.4896
	Complete	0.6923	0.5872	0.6915
	Average	**0.7691**	**0.6733**	**0.7684**
	Ward	0.7431	0.6711	0.7430
Stewart_Mature	Single	0.7548	0.7981	0.7302
	Complete	0.7346	0.5137	0.7249
	Average	**0.8802**	**0.9647**	**0.8796**
	Ward	0.6480	0.2757	0.6187
MCA_sub	Single	0.6307	0.1764	0.5780
	Complete	0.7636	0.5484	0.7631
	Average	**0.7826**	**0.5462**	**0.7846**
	Ward	0.7811	0.5425	0.7817

In this experiment, the distance measures are set to the shortest distance (Simple), the longest distance (Complete), the average distance (Average) and the Ward method (Ward), respectively, to study their impact on the performance of the algorithm. Results in Table 2 show that the distance measure in hierarchical clustering significantly affects the clustering performance. Among the six datasets tested in the experiment, the average distance metric achieves the best performance in five datasets, while the performance in the remaining Shekhar dataset ranks second, with a normalized mutual information (NMI) gap of only 0.00038 from the best-performing Ward method. Based on these results, the average distance metric is the optimal choice of distance metric in the current algorithm.

4.2 Multi-level Structural Analysis

Our analysis employs clustering visualization to showcase the significant capability of ALLSTATE in extracting meaningful insights from complex datasets. We focus on the outcomes of ALLSTATE clustering across multiple resolutions, ensuring these resolutions are in sync with the actual cell type diversity present in our dataset. As illustrated in Fig. 2a, the transition from Level 1 to Level2 resolution exemplifies the adeptness of ALLSTATE in accurately identifying specific cell types, e.g., B cells and NK cells. This precision in distinguishing cell type, avoiding the merging of unrelated cell groups, highlights the method's exceptional strength and reliability in unraveling the complex web of biological data.

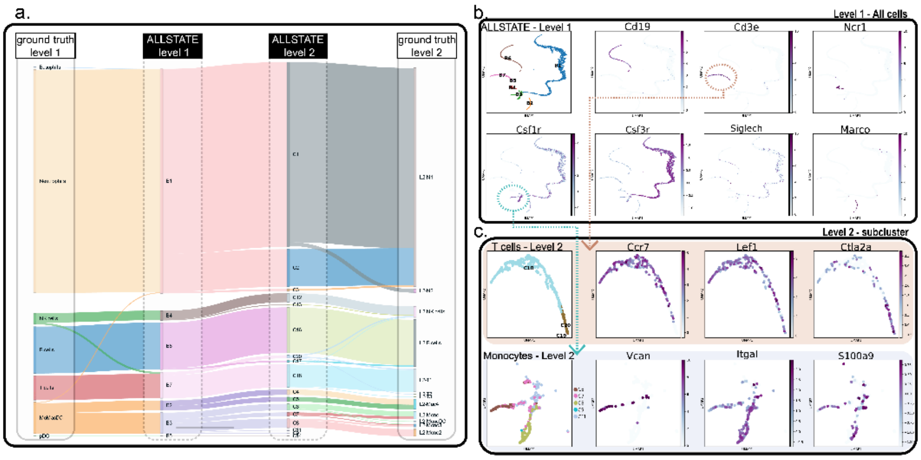


Fig. 2. Multi-level structural analysis results.

ALLSTATE allows researchers to explore the connections between cellular hierarchies and cell types at multiple levels of resolution after identifying the appropriate number of clusters. As shown in Fig. 2a, the new resolution levels from Level 1 to Level 2 provide a rich perspective for this type of analysis. This method can support tracking to the depth of each cell type branch, greatly promotes an in-depth understanding of cell subtype classification and its biological characteristics, and provides a solid foundation for exploratory cell subtype identification tasks.

To further validate the clustering performance of ALLSTATE, we extended our analysis by mapping the clustering results obtained from ALLSTATE against the expression profiles of known marker genes. As depicted in Fig. 2b, the expression patterns of these marker genes align well with clustering results of ALLSTATE at the Level 1 resolution, with each cluster corresponding to specific cell type identification markers. This alignment confirms the precision of ALLSTATE's results.

Highlighting the capacity of ALLSTATE to discern cellular nuances across multiple analytical layers, our exploration extends to a nuanced examination of subclusters within T cells and monocytes, initially identified at level 1 resolution, now analyzed at a more detailed Level 2 resolution. The findings, as illustrated in Fig. 2c, reveal that cells segregated into distinct subclusters exhibit differential expression patterns of specific marker genes. This variation in marker gene expression across subclusters not only validates ALLSTATE's ability to unravel cellular complexity at a granular level but also highlights its prowess in capturing the nuanced diversity within broader cell type categorizations.

4.3 Comparison with Existing Clustering Methods

To verify the effectiveness of ALLSTATE, we compare ALLSTATE with seven popular clustering algorithms, including PCA+HC, PCA+K-means, Louvain, Leiden, HGC, and two deep learning-based methods (scDeepCluster and DESC), on six scRNA-seq datasets. We choose adjusted rand index (ARI) [32] and normalized mutual information (NMI) [33] as the performance evaluation metrics. Higher score of these two metrics indicates better performance. The results are shown in Fig. 3.

ALLSTATE demonstrates robust clustering performance across a range of datasets and achieves comparable accuracy to established methods like Leiden, a widely accepted clustering method in single-cell cell type identification. It consistently shows high average clustering accuracy at both coarse-grained and fine-grained cell types levels. Specifically, ALLSTATE yields superior NMI values, demonstrating enhanced accuracy in matching true labels of cell types, especially noticeable in datasets GSE144735 Level 1 and Level 2, with scores of 0.87 and 0.91, respectively. This trend is mirrored in the ARI benchmarks, where our method achieves top scores, notably a 0.72 ARI on GSE144735 Level1, underscoring its robustness in identifying the correct number of clusters and their quality. The performances of multi-level also indicate ALLSTATE's robustness in handling various data complexities and its effectiveness without the need for the extensive training and parameter tuning that deep learning methods typically require.

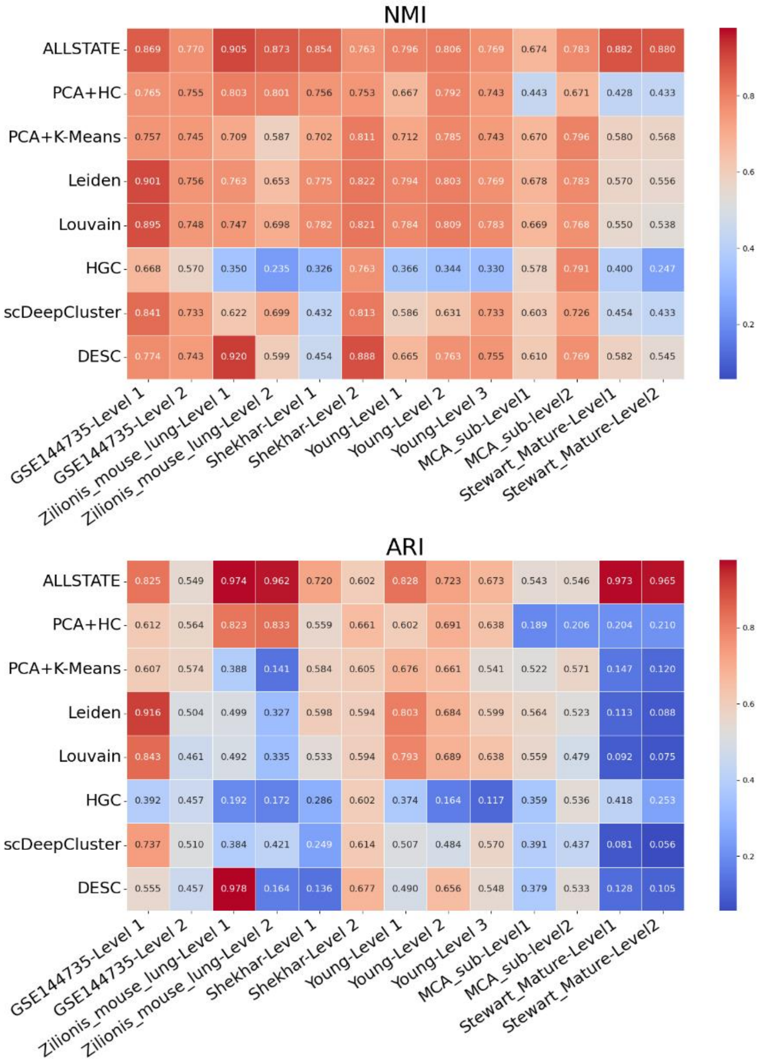


Fig. 3. Heatmap of NMI and ARI values for all methods among six datasets.

To illustrate the effectiveness of latent space representation, we use UMAP to visualize the 2D embeddings from ALLSTATE, PCA, and deep learning models as shown in Fig 4. ALLSTATE effectively delineates distinct cell populations in various datasets, such as MCA_sub, GSE144735, and Stewart_Mature, demonstrating robust clustering capabilities. In contrast, PCA and deep learning methods like scDeepCluster and DESC often create overlapping and fragmented clusters, particularly in complex datasets like Zilionis_mouse_lung. Moreover, ALLSTATE's embeddings demonstrate distinct capability in preserving the coherent and continuous distributions of cell populations across the datasets, demonstrating its exceptional ability to map complex biological continuities. Conversely, the embeddings from PCA, scDeepCluster and DESC tend to focus primarily on local structures, resulting in clusters appearing as isolated islands without clear transitional connections between related cell states.

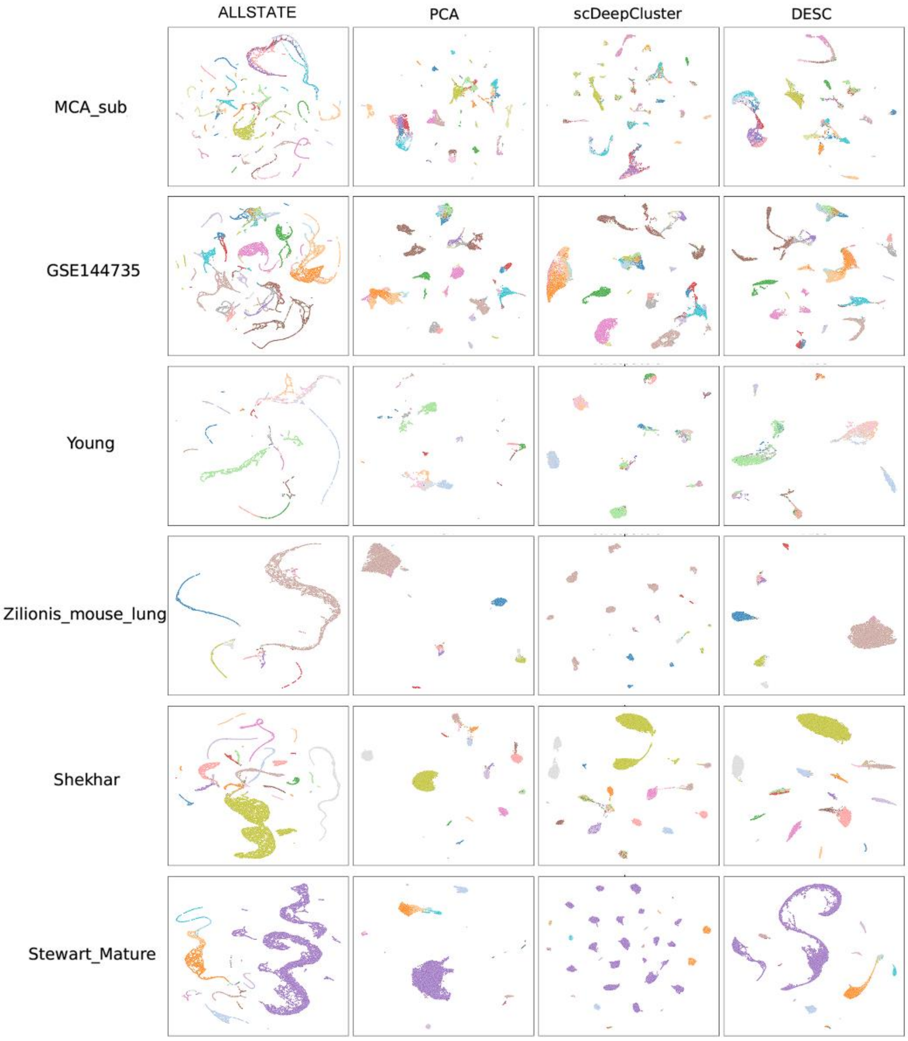


Fig. 4. Comparison of 2D visualization of embedded representations. The axes are arbitrary units. Each point represents a cell. The distinct colors of the points represent the true labels. No method uses the true label information.

Furthermore, we integrate the embedding of scDeepCluster and DESC methods into hierarchical clustering for multi-level structural analysis. We also extended our analysis by mapping the clustering results against the expression profiles of known marker genes. As shown in Fig 5, both methods have difficulty delineating subtypes of T cells within the data across multi-levels, which is as expected as deep learning-based cell clustering methods are typically trained to produce cell representations that match the cell labels provided at the training stage. As a result, the results may not preserve the fine structures existing in the distributions of interconnected cell types essential for single-cell analysis of multiple resolutions.

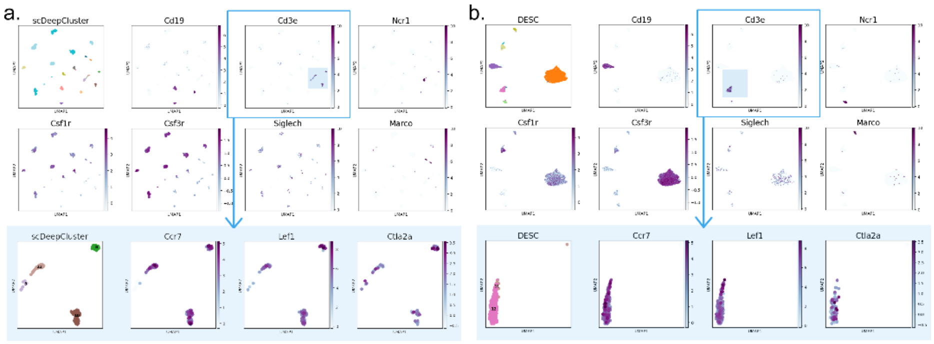


Fig. 5. Multi-level structural analysis results based on embedding of deep learning-based methods.

4.4 ALLSTATE Reveals Evolution Path among Myeloid Cell Subtypes

To further evaluate the interpretative capabilities of the ALLSTATE pipeline for downstream analysis in single-cell studies, we utilize the GSE120575 dataset from a cohort of melanoma patients undergoing PD-1 immune checkpoint blockade therapy. This dataset comprises 16,291 immune cells derived from tumor biopsies of 48 patients treated with checkpoint inhibitors. Among these individuals, 34 do not respond to the therapy while 14 are identified as responders. This dataset is chosen to rigorously assess how effectively the ALLSTATE pipeline can navigate and elucidate the intricate cellular dynamics associated with varying responses to immunotherapy.

Firstly, we identify the myeloid cells similar to the cluster of monocytes/macrophages using ALLSTATE clustering (Fig. 6a). Then we can define monocytes, macrophage and moDCs (monocyte-derived dendritic cells) by expression of CLEC10A, MARCO, APOE, CD14, CD163 and CD1C genes (Fig. 6c). By adopting PAGA, unsupervised ordering of the scRNA-sequenced myeloid cells by diffusion mapping revealed a monocytes-to-moDCs or monocytes-to-macrophages differentiation trajectory (Fig. 6d).

We also visualize the changes in gene expression of genes associated to macrophages (Fig. 6e right) or moDCs (Fig. 6e left) along the two trajectories from monocytes in pseudo-time. This reveals a progressive increase in the expression of CD163 and APOE in the early stages, followed by an upregulation of MARCO later in the macrophage trajectory. We also observed a late increase of CLEC10A along the moDCs trajectory.

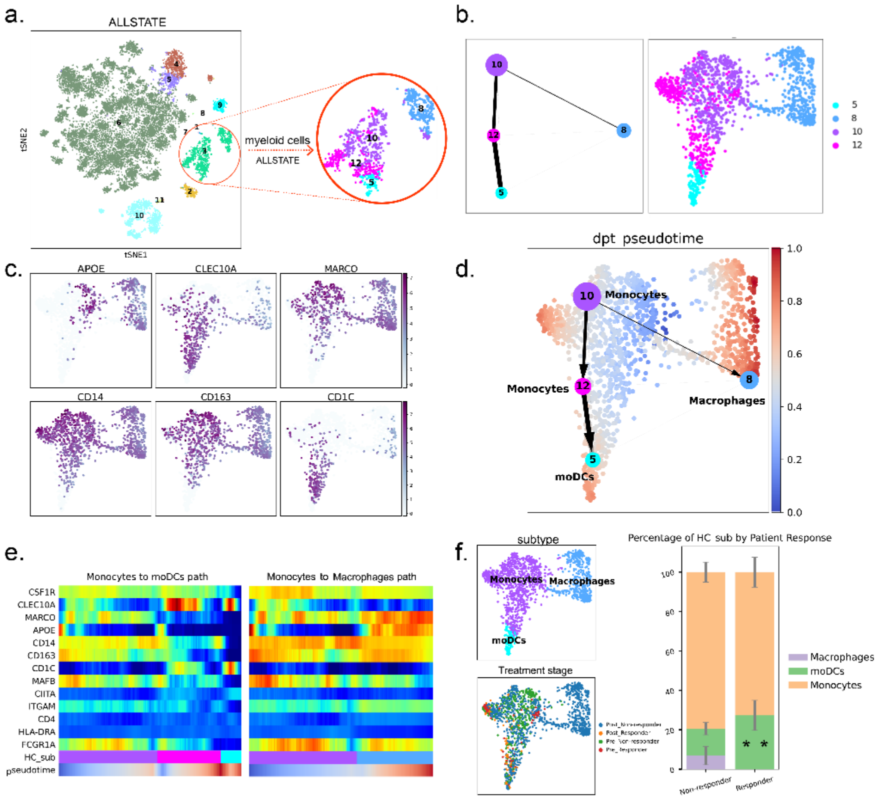


Fig. 6. Myeloid cells in human patients with melanoma show a bimodal differentiation pattern related to the therapeutic response of αPD1 therapy. (a) Single-cell RNA sequencing data [34] of tumor biopsies of patients with metastatic melanoma treated with αPD1 therapy identify myeloid cells, including four subsets generated by ALLSTATE. (b) PAGA graphs of four subsets. (c) Expression of several key genes are differentially distributed in the tumor-resident myeloid cells. (d) Using the four identified subsets as landmarks, PAGA is used to order cells in pseudo-time. (e) Quantification of monocytes, macrophages and moDCs from tumor biopsies of patients with melanoma either responding or not responding to PD1 checkpoint blockade. (f) Gene changes along PAGA paths of the differentiation process of monocytes to macrophages or dendritic cells.

Furthermore, we examine the impact of PD-1 checkpoint blockade on this cellular differentiation. By quantifying the proportionate presence of monocytes, moDCs, and macrophages, we find that moDCs are notably more prevalent in patients who exhibit a positive response to αPD1 treatment, in contrast to non-responders (Fig. 6f), suggesting that that the presence of strong moDCs infiltration in tumor microenvironment could be a predictive marker for patient response to PD1-targeted therapies.

5 Conclusion

This paper presents ALLSTATE, a scRNA-seq data analysis pipeline, which integrates non-linear dimension reduction, hierarchical clustering, PAGA topological mapping, and pseudo-time analysis to support refined deciphering of cellular diversity and differentiation paths. The experimental findings demonstrate that ALLSTATE can accurately resolution levels, and reveal complex biological characteristics and differentiation paths, thus providing a useful tool for biologists to obtain a clearer picture of underlying cell biology from data generated from their single-cell studies.

References

1. Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. 14(9), 618–630 (2013)
2. Aldridge, S., Teichmann, S.A.: Single cell transcriptomics comes of age. Nat. Commun. 11(1), 4307 (2020)
3. Zhao, T., Lyu, S., Lu, G., et al.: Sc2disease: a manually curated database of single cell transcriptome for human diseases. Nucleic Acids Res. 49(D1), D1413–D1419(2021)
4. Wang, S., Sun, S.T., Zhang, X.Y., et al.: The evolution of single-cell RNA sequencing technology and application: progress and perspectives. Int. J. Mol. Sci. 24(3),2943 (2023)
5. Zhang, A.W., O'Flanagan, C., Chavez, E.A., et al.: Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat. Methods16(10), 1007–1015 (2019)
6. Zhang, Y., Tran, D., Nguyen, T., et al.: A robust and accurate single-cell data trajectory inference method using ensemble pseudotime. BMC Bioinform. 24(1),55 (2023)
7. Wang, K., Hou, L., Wang, X., et al.: Phylovelo enhances transcriptomic velocity field mapping using monotonically expressed genes. Nat. Biotechnol. pp. 1–12(2023)
8. Sinaga, K.P., Yang, M.S.: Unsupervised K-means clustering algorithm. IEEE access8, 80716–80727 (2020)
9. Ahmed, M., Seraj, R., Islam, S.M.S.: The K-means algorithm: A comprehensive survey and performance evaluation. Electronics 9(8), 1295 (2020)
10. Chodrow, P.S., Veldt, N., Benson, A.R.: Generative hypergraph clustering: From blockmodels to modularity. Sci. Adv. 7(28), eabh1303 (2021)
11. Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9(1), 5233 (2019)
12. Tan, Y., Cahan, P.: SingleCellNet: a computational tool to classify single cell RNAseq data across platforms and across species. Cell Syst. 9(2), 207–213 (2019)
13. Aran, D., Looney, A.P., Liu, L., et al.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. 20(2),163–172 (2019)
14. Cao, Z.J., Wei, L., Lu, S., et al.: Searching large-scale scRNA-seq databases via unbiased cell embedding with cell blast. Nat. Commun. 11(1), 3458 (2020)
15. Jia, S., Lysenko, A., Boroevich, K.A., et al.: scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning. Briefings Bioinform.24(5), bbad266 (2023)
16. Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2(1), 86–97 (2012)

17. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdiscip. Rev.Comput. Stat. 2(4), 433–459 (2010)
18. Tian, T., Wan, J., Song, Q., et al.: Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat. Mach. Intell. 1(4), 191–198 (2019)
19. Li, X., Wang, K., Lyu, Y., et al.: Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat. Commun. 11(1), 2338(2020)
20. Moon, K.R., Van Dijk, D., Wang, Z., et al.: Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. 37(12), 1482–1492 (2019)
21. Wolf, F.A., Hamey, F.K., Plass, M., et al.: PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 20, 1–9 (2019)
22. Lee, H.O., Hong, Y., Etlioglu, H.E., et al.: Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat. Genet. 52(6),594–603 (2020)
23. Zilionis, R., Engblom, C., Pfirschke, C., et al.: Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. Immunity 50(5), 1317–1334 (2019)
24. Shekhar, K., Lapan, S.W., Whitney, I.E., et al.: Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166(5), 1308–1323(2016)
25. Young, M.D., Mitchell, T.J., Vieira Braga, F.A., et al.: Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science 361(6402),594–599 (2018)
26. Stewart, B.J., Ferdinand, J.R., Young, M.D., et al.: Spatiotemporal immune zonation of the human kidney. Science 365(6460), 1461–1466(2019)
27. Fei, L., Chen, H., Ma, L., et al.: Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development. Nat. Genet. 54(7), 1051–1061 (2022)
28. Ma, S., Dai, Y.: Principal component analysis based methods in bioinformatics studies. Briefings Bioinform. 12(6), 714–722 (2011)
29. Xu, Q., Ding, C., Liu, J., Luo, B.: PCA-guided search for K-means. Pattern Recogn. Lett. 54, 50–55 (2015)
30. Que, X., Checconi, F., Petrini, F., Gunnels, J.A.: Scalable community detection with the Louvain algorithm. In: IEEE Int. Parallel Distributed Process. Symp. (IPDPS). pp. 28–37. IEEE (2015)
31. Zou, Z., Hua, K., Zhang, X.: HGC: fast hierarchical clustering for large-scale single cell data. Bioinformatics 37(21), 3964–3965 (2021)
32. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification 2, 193–218(1985)
33. McDaid, A.F., Greene, D., Hurley, N.: Normalized mutual information to evaluate overlapping community finding algorithms. arXiv preprint arXiv:1110.2515 (2011)
34. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., et al.: Defining T cell states associated with response to checkpoint immunotherapy in melanoma. Cell 175(4),998–1013 (2018)