Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/57/20241249

VisRNA: Interactive web server for scRNA-seq data analysis to discover therapeutic targets for non-small cell lung cancer

Xingchuan Ma

Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic, 44195, Cleveland, OH, USA

xingchuan_ma@portsmouthabbey.org

Abstract. Motivation: Non-small cell lung cancer (NSCLC) is a leading cause of lung cancer diagnoses and mortality worldwide. Single-cell RNA sequencing (scRNA-seq) has transformed our molecular understanding of cancer by revealing gene dysregulation in individual cells. This granularity allows the discovery of new therapeutic targets, promising NSCLC treatment strategies. Method: The study introduces VisRNA, a Python-based web application for analyzing scRNA-seq data statistically and functionally. VisRNA uses advanced machine learning algorithms to reduce multiple dimensionalities in scRNA-seq data and automatically annotate cell types. This method allows gene expression analysis across NSCLC tumor cell populations. Results: VisRNA analysis of NSCLC scRNA-seq datasets identified 14 distinct cell types with distinct gene expression patterns. VisRNA found several significantly differentially expressed genes in these cell populations that could be therapeutic targets. The application ranked drug candidates by molecular docking simulation performance, indicating potential efficacy against targets. Conclusion: VisRNA is a powerful tool for identifying new therapeutic targets and drug candidates for NSCLC, highlighting a significant advancement in scRNA-seq data analysis. VisRNA helps understand the disease's molecular basis by examining gene expression in individual NSCLC tumor cells, which may lead to better treatments.

Keywords: scRNA-seq, non-small cell lung cancer, visualization platform

1. Introduction

Lung cancer is the second most prevalent and the most lethal type of cancer [1-3]. Specifically, non-small lung cancer (NSCLC) accounts for more than 87% of lung cancer cases. NSCLC begins at the cellular level, where abnormal cells reproduce rapidly. Genetic factors play a crucial role in understanding the mechanism of NSCLC and the prognosis of NSCLC [4-6]. Most NSCLCs are driven by chromosomal instability (CIN), providing genetic diversity to promote cancer progression. These unique attributes collectively contribute to the high complexity and heterogeneity of the cancer genetic landscape [7].

Single-cell RNA sequencing (scRNA-seq) emerges as a high-resolution method for identifying gene expression profiles across different cell types. scRNA-seq examine the gene expression levels in an individual cell by measuring the mRNA expressions. The transcriptomic profile of each cell type differs from the others and exhibits heterogeneity in the tumor microenvironment. This approach is suitable for investigating subcellular-level biology or immune cell heterogeneity [8-9], helping investigators

DOI: 10.54254/2753-8818/57/20241249

determine marker genes more efficiently. Consequently, effective visualization and downstream analysis of scRNA-seq data are pivotal for uncovering cancer biomarkers and potential therapeutic targets.

In response to the rapid evolution of computational tools integrating scRNA-seq data of different cancer types and visualization, numerous software applications have emerged for downstream analysis of scRNA-seq. Recently, Zeng et al. developed the CancerSCEM [10-11] database that integrates datasets from scRNA sequencing over the past decade. This dataset encompasses a diverse array of cancers and their sequencing data, offering a user-friendly interface for researchers. Such tools are instrumental in unraveling critical genes indicative of tumor cell heterogeneity. However, only a few studies have focused on finding potential therapeutic targets that may be used for drug repurposing after identifying the marker gene for each cell type. Beyondcell addresses a significant cancer treatment challenge of tumor cell heterogeneity. Variation within tumors often leads to cell responses to treatment options that lead to drug resistance and therapeutic failures. Beyondcell identifies cell subpopulations based on their drug responses to overcome this challenge [12].

I designed an end-to-end web application to perform downstream analysis of scRNA-seq data for the semiautomatic discovery of therapeutic targets of NSCLC in this project. I applied the web app for the drug repurposing of NSCLC. I first performed dimensionality reduction and visualization on NSCLC scRNA sequencing data, followed by cell type annotation using a machine learning approach, differentially expressed gene (DEG) analysis, and functional enrichment analysis. The 10 high-ranking DEGs were identified for each cell type and mapped to drug targets to examine the potential binding affinity with drug candidates of NSCLC. The whole pipeline was hosted on a public web server (VisRNA) with a user-friendly interface, which will promote high-throughput biomarker discovery based on scRNA-seq data and assist drug discovery for NSCLC.

2. Methods

2.1. Dataset

The dataset used in the project was acquired from a publicly accessible database called CancerSCEM [10-11], which contains processed scRNA-seq data in the format of matrices. Specifically, I obtained data from GSE123904 (GEO accession number) with 14 tumor and adjacent standard samples, which provided the gene expression matrices, cell components, differential expression genes, and some critical molecules with the cell interactions. The normalized gene expression matrix of the NSCLC data was extracted for the study.

2.2. Downstream analysis of scRNA-seq data

2.2.1. Visualizing data in low-dimensional space. Three dimensionality reduction methods were implemented on VisRNA: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). PCA projects high-dimensional scRNA-seq data into a linearly orthogonal low-dimensional vector space [13-14]. t-SNE converts cell similarities into probability and includes information from cell clusters into visualization by redefining the likelihood. It computes spatial cell maps in low dimensions by minimizing Kullback-Leibler divergence [15]. UMAP creates a high-dimensional graph representation of the data before constructing a low-dimensional graph that is as structurally comparable as feasible. UMAP is well known for its computational efficiency and astounding ability to preserve the global structure of the data itself [16-17].

2.2.2. Clustering cells into putative subpopulations. After dimension reduction, cells with similar gene expressions would be close to each other on the plot and vice versa. Clustering analysis was performed to identify subpopulations of cells.

DOI: 10.54254/2753-8818/57/20241249

For PCA and t-SNE, K-means clustering was performed. For UMAP, Leiden clustering [18] was adopted to perform clustering analysis for subsequent cell type annotation. Several critical parameters that may determine the number of clusters are customizable.

2.2.3. Cell type annotation. CellTypist (https://www.celltypist.org/) was used to perform cell type annotation, a machine learning-based approach for rapid and accurate cell type recognition created to resolve immune cell heterogeneity across tissues [19]. CellTypist contains a few pretrained models for different tissue types, such as heart, lung, kidney, and organism types, including humans and mice. If some clusters could not be annotated, CellMarker (http://xteam.xbio.top/CellMarker/) was used to find the possible appropriate marker gene. As a result, based on DEGs and the CellMarker database, several vital genes overlap in both datasets to determine the cell type for ambiguous clusters [20].

2.2.4. Differentially expressed gene analysis between clusters and cell types. A Welch t-test was performed on the log-expression value for each gene and each pair of clusters to find the marker gene between different cell clusters. The goal is to determine differentially expressed genes (DEGs) by comparing them to the other cells in the cluster. The top DEGs are also good candidates for markers, since they distinguish themselves from clusters. The results of the DEGs are summarized in a table that directly compares the DEGs of each cluster. A heat map was used to show DEGs to visualize gene expression. High-ranking DEGs possess robust and constant up or down-regulation in one of the clusters compared to others.

2.3. Availability and Implementation

The complete source code is available on GitHub (https://github.com/ryanmxc/VisRNA). It can be downloaded as a package for offline usage. Additional information on the methods used, including data acquisition, downstream analysis of scRNA-seq data, cell type annotation, differentially expressed gene analysis, gene enrichment analysis, and interactive visualization web server development, is available on the Github repository.

3. Results and Discussion

3.1. Framework of VisRNA for scRNA-seq data visualization

The visualization of scRNA data concerning NSCLC encompasses several essential steps. VisRNA can be hosted locally. All the essential modules for processing and visualization are seamlessly integrated into the VisRNA web server, powered by Python and Streamlit (**Figure 1**). The initial steps involve the extraction of scRNA-seq-processed data from the CancerSCEM database. Subsequently, data dimensionality is reduced through PCA, t-SNE, and UMAP techniques. Then, K-means and Leiden clustering were used to distinguish cell clustering. Furthermore, a hierarchical progressive machine learning model, Cell Typist, with its comprehensive atlas of gene sequences, was utilized to achieve accurate cell-type predictions. This model can also discern marker genes through differential gene expression (DEG) analysis. These marker genes serve as critical indicators for early disease suppression. This platform also generates a heat diagram showing differentially expressed genes in each cell type. **Figure 2** shows screenshots of the essential steps in the VisRNA web application, including data upload, dimensionality reduction, cell type annotation, and DEG analysis. Users can effortlessly upload preprocessed scRNA-seq data in CSV format, entrusting the platform to execute automatic processing. Furthermore, all clustering diagrams, cell type annotations, and DEG analysis results are downloadable for further analysis, including drug–target interaction prediction.

DOI: 10.54254/2753-8818/57/20241249

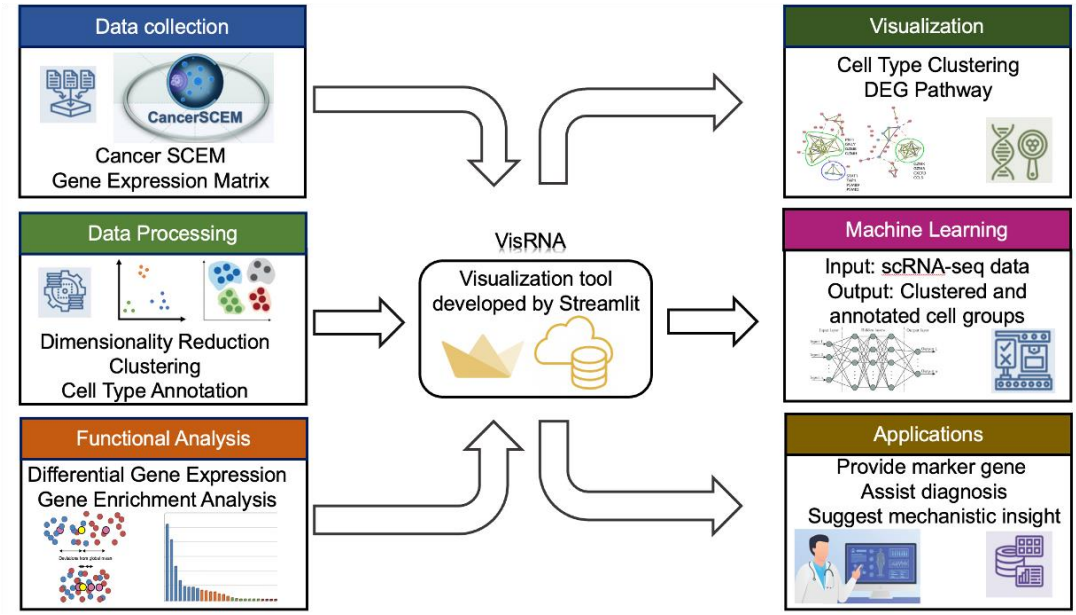

Figure 1. Framework of VisRNA for scRNA-seq data visualization of NSCLC.

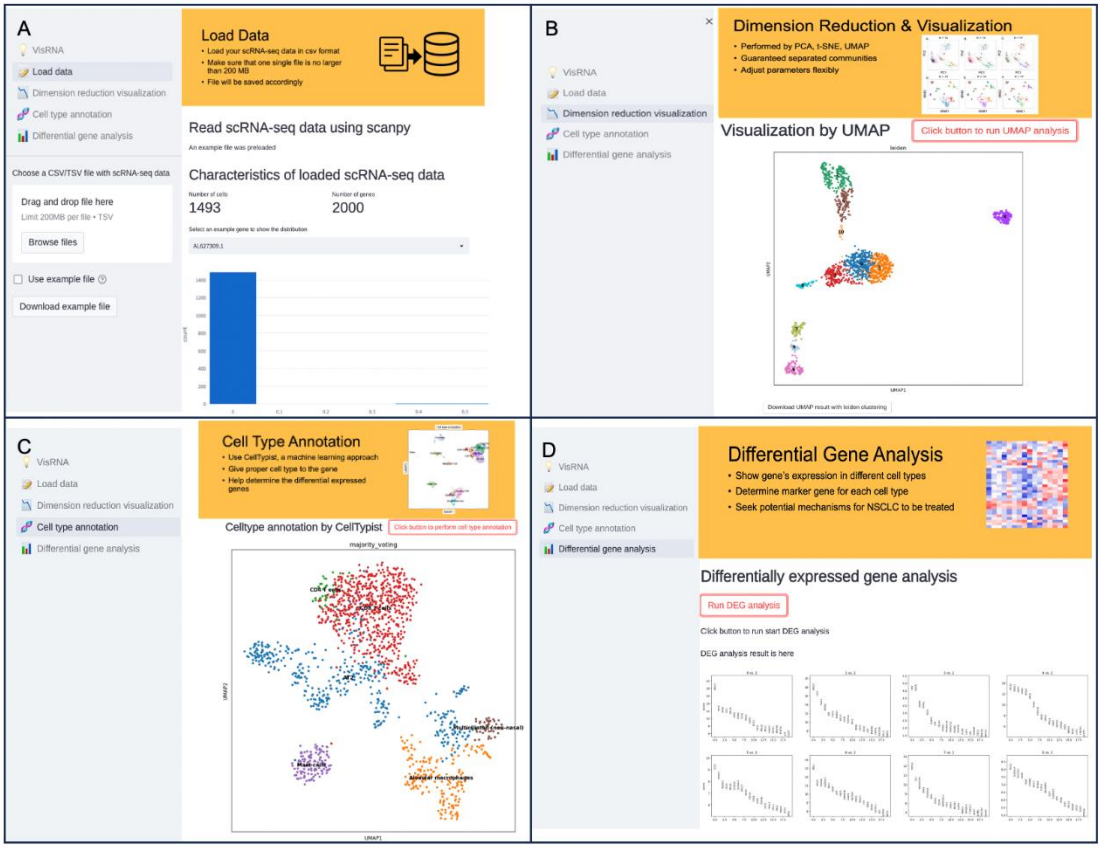

Figure 2. Screenshot of VisRNA for scRNA-seq data visualization. Panel A shows the page of load processed scRNA-seq data. Panel B shows the dimensionality reduction page, and the figure can be downloaded using the download button. Panel C shows the cell type annotation done by CellTypist. Panel D shows the differential gene analysis on the webpage.

DOI: 10.54254/2753-8818/57/20241249

3.2. Dimensionality reduction and clustering of scRNA-seq data for NSCLC

I employed three dimensionality reduction techniques and different clustering methods in the dimensionality reduction module of VisRNA, as described in the Methods section. Users can fine-tune parameters to determine cluster counts for subsequent cell-type annotations. This study compared PCA, t-SNE, and UMAP techniques for NSCLC scRNA-seq data visualization. The number of clusters ranged from 12 to 17. **Figure 3** demonstrated that t-SNE and UMAP plots formed more concentrated clusters than PCA. Leiden clustering was deployed with UMAP generating the most discrete clusters, a desirable outcome for cell-type annotation.

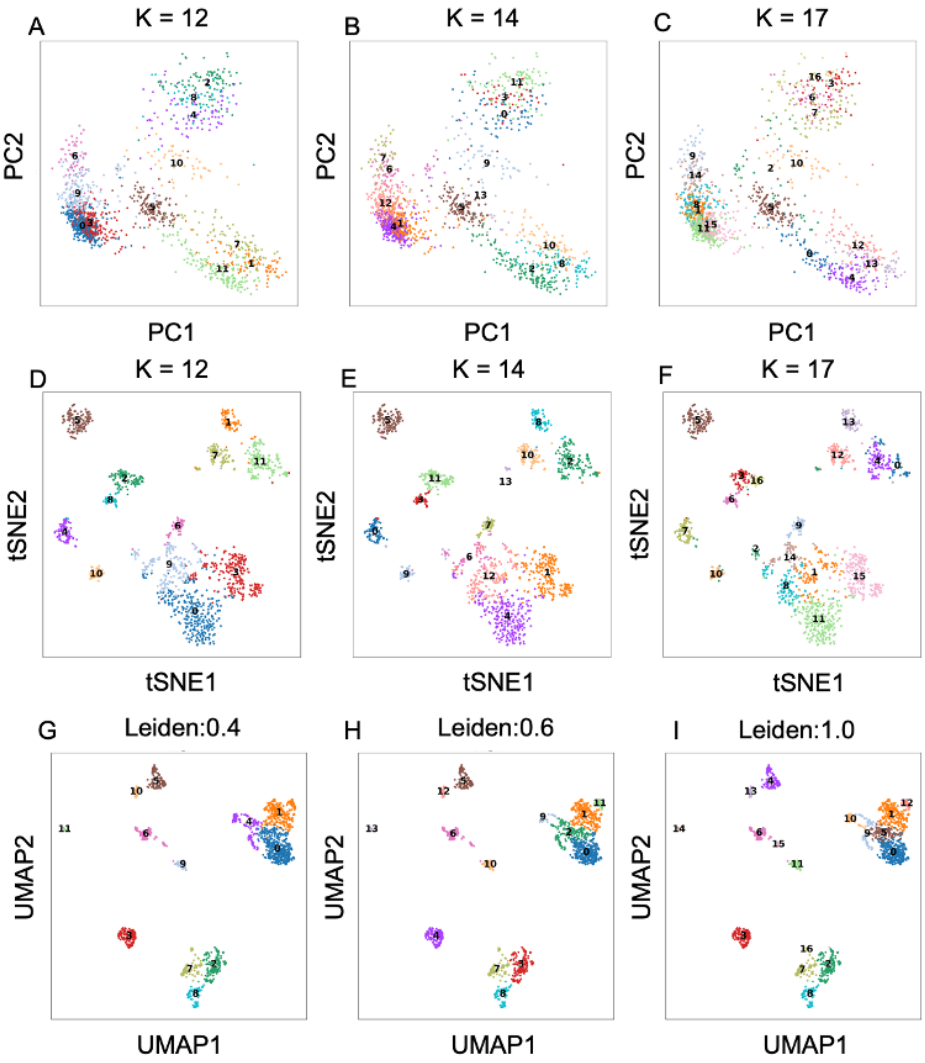

Figure 3. Dimensionality reduction and clustering by PCA, t-SNE and UMAP. Panel A-C shows the dimensionality reduction by PCA and the clusters are divided by K-means into 12, 14, or 17. Panel D-F shows the dimensionality reduction by t-SNE and the clusters are determined by K-means into 12, 14, or 17. Panel G-I shows the dimensionality reduction by UMAP and the clusters are divided by Leiden clustering into 12, 14, or 17.

3.3. Cell type annotation and differentially expressed gene analysis

After performing cell type annotation using CellTypist (https://www.celltypist.org/), I utilized known marker genes to visualize gene expression across all cells to validate the reliability of cell type annotation, as shown in **Figure 4**. The marker genes for mast cells and fibroblasts were distinctly located within

DOI: 10.54254/2753-8818/57/20241249

their respective clusters, allowing for the confident annotation of these cell types. These cells serve as prime examples of explicit annotation, which helps eliminate ambiguity during the cell-type annotation process. However, certain cell types, such as CD8 Tem and CD8 naive T cells, exhibit similar expression profiles, thus increasing annotation uncertainty. Nevertheless, the putative annotations for these cell types allow for further differentially expressed gene analysis. Each cell type has a few differentially expressed genes. Eventually, 17 clusters were annotated with 14 cell types, with some of the clusters identified as the same.

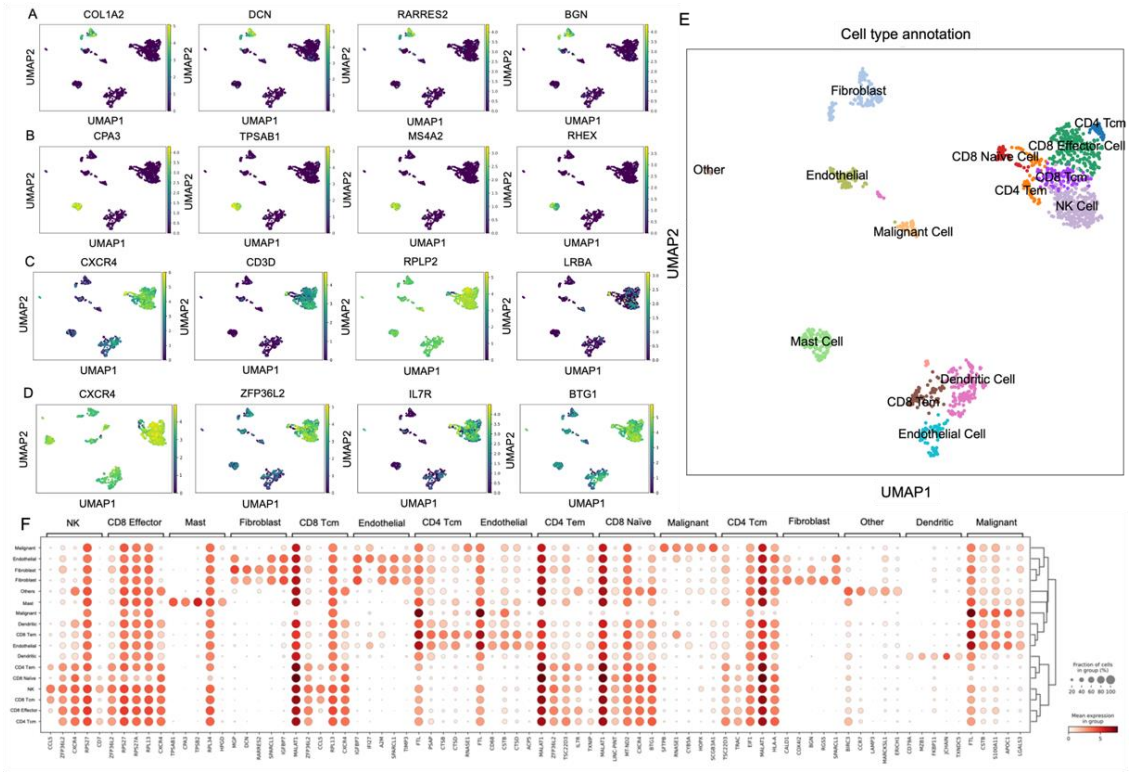

Figure 4. Cell type annotation and differentially expressed gene analysis of NSCLC data. Panel A-D show the feature plot of the most differentially expressed gene in mast cells, fibroblasts, CD8 Tem and CD8 naive T cells in order. Panel E shows the cell type annotation result performed by CellTypist. Panel F shows the differentially expressed gene analysis with annotated cell type.

4. Conclusion

VisRNA harmoniously integrates the procedures for downloading scRNA-seq data, including dimensionality reduction, clustering, cell type annotation, and differential gene analysis, producing a user-friendly interface in which all the data were transformed into visualizations for biological interpretations to find therapeutic targets. The machine learning–based method, cell type annotation, is seamlessly embedded in VisRNA, allowing for rapid identification of prospective therapeutic targets. Differential gene analysis enables rapid identification of potential therapeutic targets.VisRNA is an interactive scRNA-seq data visualization platform for efficient data analysis and therapeutic target discovery.

Acknowledgment

I thank the high-performance computing service provided by National Institute of Health Data Science of China, Shandong University.

References

[1] Lung Cancer Statistics | How Common is Lung Cancer? https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html.
[2] Evison, M. & AstraZeneca UK Limited. The current treatment landscape in the UK for stage III NSCLC. Br J Cancer 123, 3–9 (2020).
[3] Duma, N., Santana-Davila, R. & Molina, J. R. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. Mayo Clin Proc 94, 1623–1640 (2019).
[4] Non-Small Cell Lung Cancer. Yale Medicine https://www.yalemedicine.org/conditions/non-small-cell-lung-cancer.
[5] Jonna, S. & Subramaniam, D. S. Molecular diagnostics and targeted therapies in non-small cell lung cancer (NSCLC): an update. Discov Med 27, 167–170 (2019).
[6] Osmani, L., Askin, F., Gabrielson, E. & Li, Q. K. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. Semin Cancer Biol 52, 103–109 (2018).
[7] Monteverde, T. et al. CKAP2L Promotes Non-Small Cell Lung Cancer Progression through Regulation of Transcription Elongation. Cancer Res 81, 1719–1731 (2021).
[8] He, S. et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat Biotechnol 1–13 (2022) doi:10.1038/s41587-022-01483-z.
[9] Vallejo, J., Cochain, C., Zernecke, A. & Ley, K. Heterogeneity of immune cells in human atherosclerosis revealed by scRNA-Seq. Cardiovasc Res 117, 2537–2543 (2021).
[10] Zeng, J. et al. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res 50, D1147–D1155 (2022).
[11] Downloads - Cancer Single-cell Expression Map - National Genomics Data Center - CNCB-NGDC. https://ngdc.cncb.ac.cn/cancerscem/downloads.
[12] Fustero-Torre, C. et al. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. Genome Medicine 13, 187 (2021).
[13] Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biology 20, 269 (2019).
[14] Liu, Z. Visualizing Single-Cell RNA-seq Data with Semisupervised Principal Component Analysis. International Journal of Molecular Sciences 21, (2020).
[15] Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. Nat Commun 10, 5416 (2019).
[16] Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 37, 38–44 (2019).
[17] Patel, S. Guide to Dimensionality Reduction in single cell RNA-seq analysis. Medium https://towardsdatascience.com/guide-to-dimensionality-reduction-in-single-cell-rna-seq-analysis-1d77284eed1c (2020).
[18] Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019).
[19] Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science 376, eabl5197 (2022).
[20] Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Research 47, D721–D728 (2019).
[21] UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51, D523–D531 (2022).
[22] Wang, Y.-Y. et al. CeDR Atlas: a knowledgebase of cellular drug response. Nucleic Acids Research 50, D1164–D1171 (2022).
[23] Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem 31, 455–461 (2010).

DOI: 10.54254/2753-8818/57/20241249

[24] Chang, S.-H. et al. Dihydroergotamine Tartrate Induces Lung Cancer Cell Death through Apoptosis and Mitophagy. Chemotherapy 61, 304–312 (2016).
[25] Chen, H. et al. Sulfonylurea receptor 1-expressing cancer cells induce cancer-associated fibroblasts to promote non-small cell lung cancer progression. Cancer Lett 536, 215611 (2022).
[26] Zhang, S. et al. Anticancer effects of ikarugamycin and astemizole identified in a screen for stimulators of cellular immune responses. J Immunother Cancer 11, e006785 (2023).