2023 S.T. Yau High School Science Award

Research Report

The Team

Name of team member: Xingchuan Ma
School: Portsmouth Abbey School
City, Country: Portsmouth, RI

Name of supervising teacher: Stephen Zins
Job Title: Head of Science Department, Biology Teacher
School/Institution: Portsmouth Abbey School
City, Country: Portsmouth, RI

Title of Research Report

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

Date

08/16/2023

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

Xingchuan Ma

Abstract

Motivation: Non-small cell lung cancer (NSCLC) is a main category of lung cancer and leading cause of death worldwide. The use of single-cell RNA-sequencing (scRNA-seq) can detect the disrupted genes and mechanisms at the single-cell level, thereby contributing to the discovery of therapeutic target of NSCLC.

Results: This study developed an efficient python-based visualization web application, called VisRNA, for statistical and functional analysis using scRNA-seq data. VisRNA performed multiple dimensionality reduction techniques of scRNA-seq data, followed by automatic cell type annotation by machine learning. By analyzing the differentially expressed genes among different cell clusters, VisRNA identified the putative therapeutic target and drug candidates, which were ranked by molecular docking simulation results. As a proof of concept, VisRNA integrated the scRNA-seq data of NSCLC, and identified 14 cell types with differentially expressed genes. The potential drug, dihydroergotamine, showed the lowest binding affinity for different targets among multiple cell types, indicating potential therapeutic target and drug candidates for NSCLC.

Availability and Implementation: VisRNA is freely available on the web (https://visrna.streamlit.app/), supporting interactive data uploading, scRNA-seq data downstream processing and visualization, and functional analysis. The platform is implemented in Python and Streamlit and supports all major browsers. The source code is freely available via github (https://github.com/ryanmxc/VisRNA).

Supplementary Information: 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 VisRNA website.

Keywords: NSCLC, scRNA-seq, marker gene, drug targets, drug candidates

Declaration of Academic Integrity

The participating team declares that the paper submitted is comprised of original research and results obtained under the guidance of the instructor. To the team's best knowledge, the paper does not contain research results, published or not, from a person who is not a team member, except for the content listed in the references and the acknowledgment. If there is any misinformation, we are willing to take all the related responsibilities.

Names of team members: Xingchuan Ma

Signatures of team members: 

Name of the instructor: Stephen Zins

Signature of the instructor: 

Date: 08/06/2023

Table of Contents

1. Introduction

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

Single-cell RNA sequencing (scRNA-seq) offers a high-resolution method to identify the gene expression profiles across different types of cells, which examines the gene expression levels in an individual cell by measuring the mRNA expressions. The transcriptomic profile of each cell type differs from the other and exhibits the heterogeneity in the tumor microenvironment. This technique is suitable when investigating subcellular level biology or immune cell heterogeneity[8-9] which help investigators figure out marker genes more efficiently. Therefore, visualization and downstream analysis of scRNA-seq data is critical for cancer biomarker and therapeutic target discovery.

With the fast development of computational tools that integrates scRNA-seq data of different cancer types and visualization, a variety of software regarding downstream analysis of scRNA-seq data have been widely developed. Recently, Zeng et al. developed a database that integrates the scRNA-seq datasets over the decade called CancerSCEM[10-11]. This dataset focuses on integrating a variety of cancers with the sequencing data and forms a user-friendly interface for other investigators to use. These tools facilitate the identification of key genes that may imply the tumor cell heterogeneity. However, after identifying the marker gene for each cell type, only a few studies focus on finding potential therapeutic targets that may be used for drug repurposing. Beyondcell addresses a major 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].

In this project, we designed an end-to-end web application to perform downstream analysis for scRNA-seq data for the semi-automatic discovery of therapeutic target of NSCLC and applied the web app for drug repurposing of NSCLC. We first performed dimensionality reduction and visualization on the scRNA-seq data of NSCLC, followed by cell type annotation via a machine learning approach, differentially expressed gene (DEG) analysis, and functional enrichment analysis. The top ten DEGs were identified for each cell type and were mapped onto 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 the high-throughput biomarker discovery based on scRNA-seq data and assists 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 a processed the scRNA-seq data in format of matrices. Specifically, we obtained the data from GSE123904 (GEO accession number) with 14 tumor and adjacent normal samples, which provides the gene expression matrices, cell components, differential expression genes, and some critical molecules with the cell interactions. The normalized gene expression matrix of NSCLC data was extracted to conduct 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, including principal component analysis (PCA), t-distributed stochastic neighbor

embedding(t-SNE), and Uniform Manifold Approximation and Projection (UMAP). PCA projects the high-dimensional scRNA-seq data into the 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 the 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 the 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 are close to each other on the plot and vice versa. Clustering analysis is performed to identify the subpopulations of the cells.

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

2.2.3. Cell type annotation

CellTypist (https://www.celltypist.org/) was used to perform the cell type annotation, which is a machine learning-based approach for rapid and accurate cell type recognition and created to resolve immune cell heterogeneity across tissues[19]. CellTypist contains a few pretrained models for different tissue types, such as heart, lung, and kidney, and organism types, including human and mouse. In case some clusters cannot be annotated, CellMarker (http://xteam.xbio.top/CellMarker/) was used to find the possible appropriate marker gene. As a result, based on the differentially expressed genes and CellMarker database, several key genes overlapping in both dataset to determine the cell type for the ambiguous clusters[20].

2.2.4. Differentially expressed gene analysis between clusters and cell type

To find the marker gene between different cell clusters, the Welch t-test is performed on the log-expression value for each gene and each pair of clusters. The goal is to figure out the 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 compared to clusters. The results of DEGs are summarized in a table that directly compares the DEGs of each cluster. The heatmap is used to show the DEGs to visualize the gene expression. Top DEGs possess robust and constant up or down-regulation in one of the clusters compared to other clusters.

2.3. Drug-target interaction by molecular docking

We selected the top ten marker genes from each cell type and used UniProt (https://www.uniprot.org/) to map those genes to their corresponding proteins [21]. These proteins were regarded as the potential drug targets, which were subjected to molecular docking in next stage. To generate drug candidates, we used CeDR[22] (https://ngdc.cncb.ac.cn/cedr/home) and selected the top drug candidates from multiple datasets (DR000345, DR000346, DR000349, DR000350, DR000352, DR000355), which generated 46 drug candidates. Finally, each protein target and drug candidate were docked by Autodock Vina and the interaction were evaluated by the binding affinities. The docking processes were run 10 times with random seeds and the conformation with lowest binding affinity was selected as the representing structure[23].

3. Results and Discussion

3.1. Framework of VisRNA for scRNA-seq data visualization

The framework for visualizing single-cell RNA (scRNA) data related to non-small cell lung cancer (NSCLC) involves several essential steps. VisRNA can be hosted locally or accessed remotely via a cloud-based server (https://visrna.streamlit.app/). All the essential modules for processing and visualization are integrated into the

VisRNA web server developed with Python and Streamlit (**Figure 1**). First, we extracted the scRNA-seq processed data from the CancerSCEM database. The data is processed by PCA, t-SNE, and UMAP in dimension reduction. Then, K-means and Leiden clustering are used to distinguish cell clustering. Afterwards, a hierarchical progressive machine learning model, Cell Typist, with its comprehensive atlas of gene sequences, is utilized to achieve accurate cell-type predictions. The model can also identify 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 showcasing differentially expressed genes in each cell type. **Figure 2** shows the screenshots of the essential steps in VisRNA web application, including data upload, dimensionality reduction, cell type annotation, and differential gene expression analysis. Users can easily upload preprocessed scRNA-seq data in CSV format and have it automatically processed by the platform. In addition, all clustering diagrams and cell type annotation, result of DEG analysis are downloadable for further analysis such as drug-target interaction prediction.

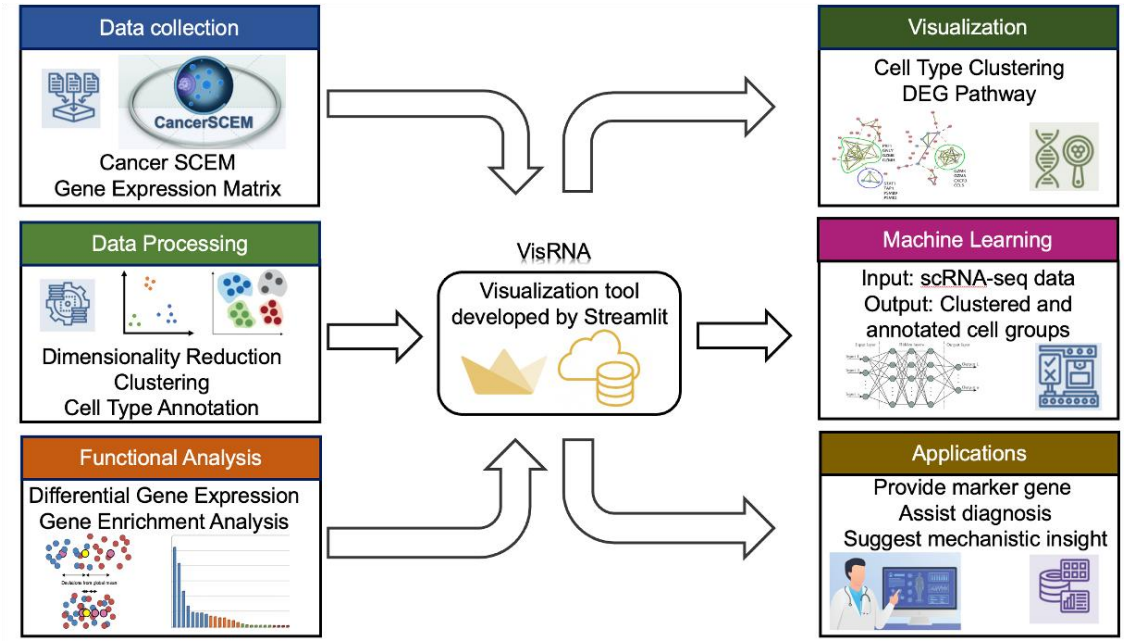

Figure 1. Framework of VisRNA for scRNA-seq data visualization of NSCLC.

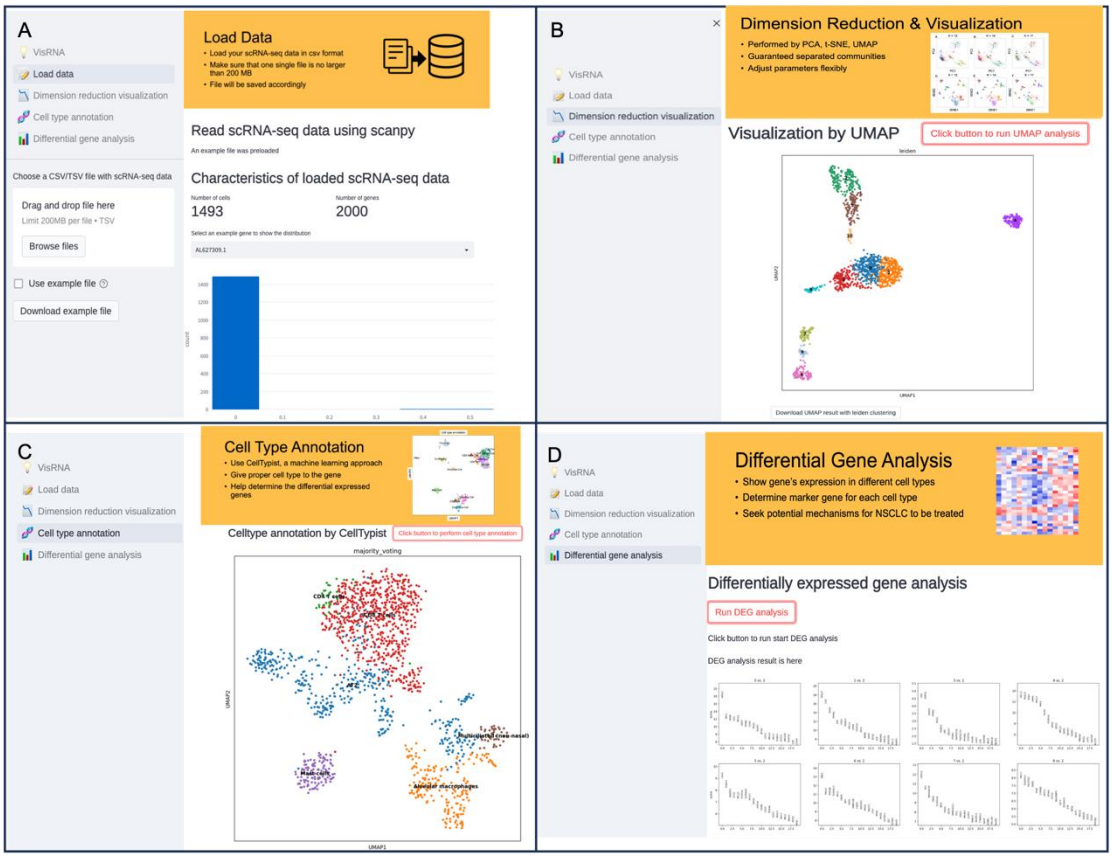

Figure 2. Screenshot of VisRNA for scRNA-seq data visualization. Panel A shows the page of load processed scRNA-seq data. Panel B shows the dimensionality reduction page, and the figure can be downloaded using the download button. Panel C shows the cell type annotation done by CellTypist. Panel D shows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

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

In the dimensionality reduction module of VisRNA, we employed three dimensionality reduction techniques and different clustering methods as described in Methods section. The users can tune the parameters to define the number of clusters for subsequent cell-type annotation. In this study, we compared PCA, t-SNE, and UMAP techniques for NSCLC scRNA-seq data visualization. The number of clusters ranges from 12 to 17. As shown in **Figure 3**, compared with PCA, t-SNE and UMAP plots form more concentrated clusters. Leiden clustering was deployed and UMAP provided the most discrete clusters, which was desirable for cell type annotation.

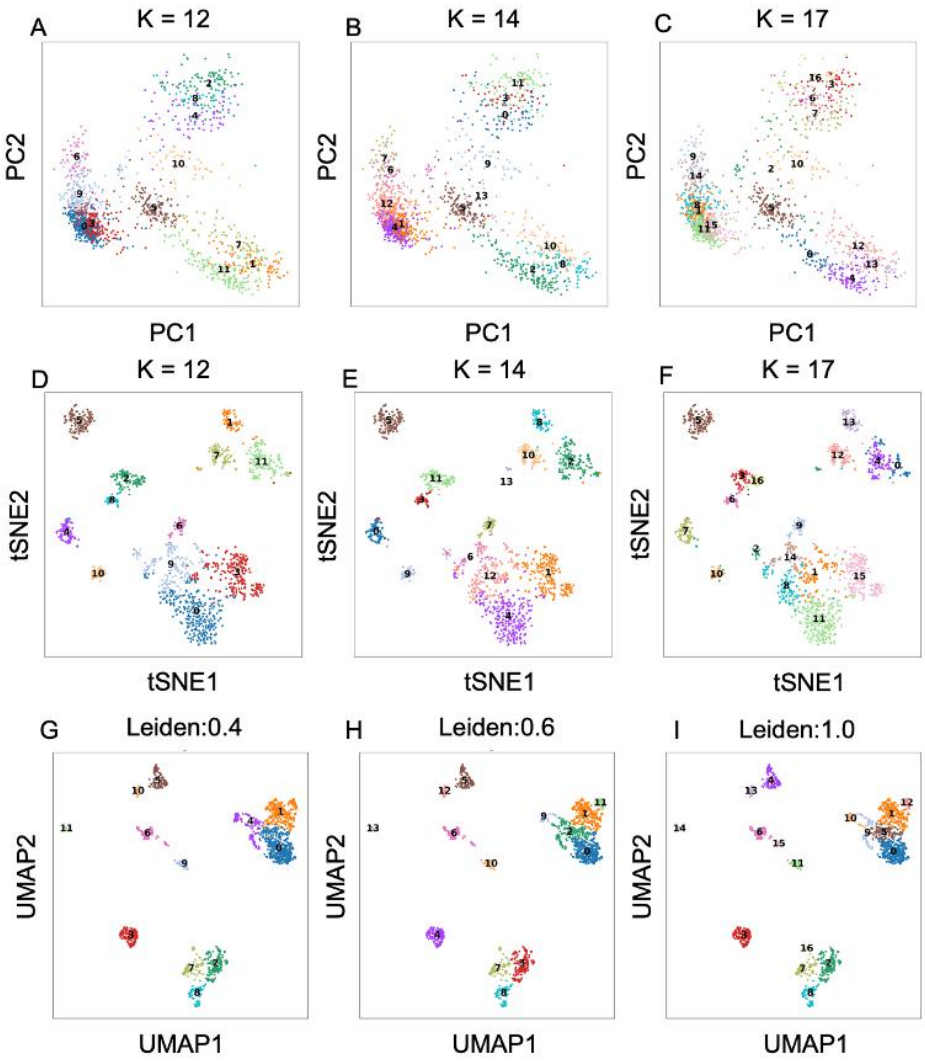

Figure 3. Dimensionality reduction and clustering by PCA, t-SNE and UMAP. Panel A-C shows the dimensionality reduction by PCA and the clusters are divided by K-means into 12, 14, or 17. Panel D-F shows the dimensionality reduction by t-SNE and the clusters are determined by K-means into 12, 14, or 17. 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 by CellTypist (https://www.celltypist.org/), we utilized known marker genes to visualize the gene expression across all cells to validate the reliability of cell type annotation as shown in **Figure 4**. Marker genes for mast cells and fibroblasts were distinctly located within their respective clusters, allowing for confident annotation of these cell types. These cells serve as prime

examples of explicit annotation, which aids in eliminating ambiguity during the cell type annotation process. However, certain cell types, such as CD8 Tem and CD8 naive T cells, exhibit similar expression profiles, increasing the annotation uncertainty. Nevertheless, the putative annotations for these cell types allows further differentially expressed gene analysis. Each cell type has a few differentially expressed genes. Eventually, 17 clusters were annotated with 14 cell type with some of the clusters identified as the same one.

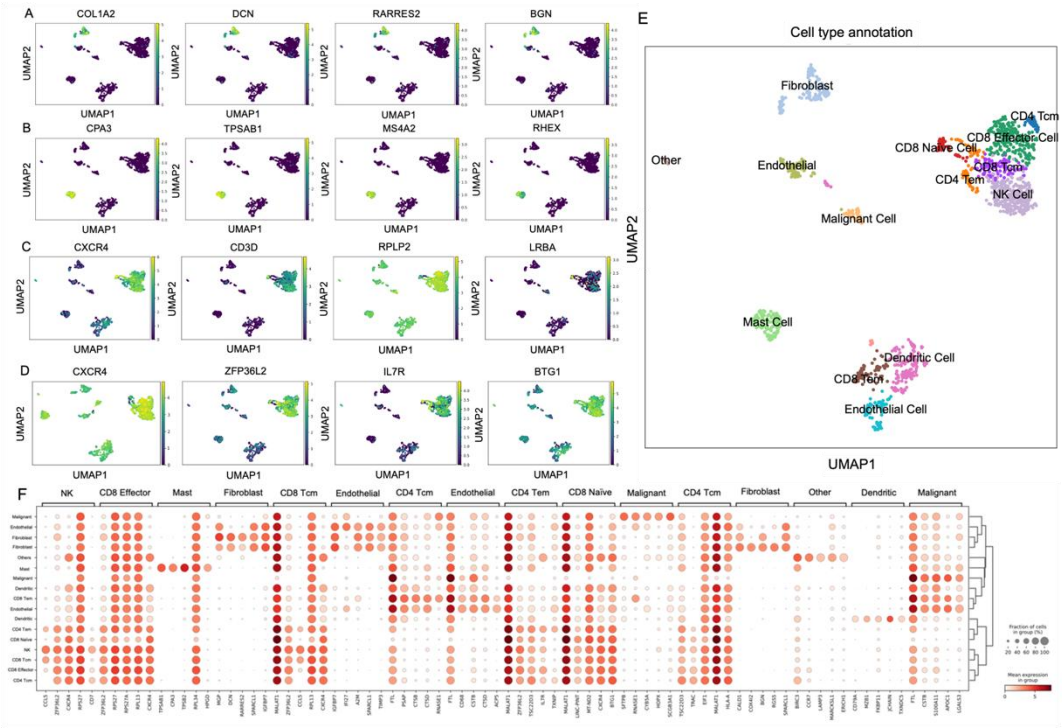

Figure 4. Cell type annotation and differentially expressed gene analysis of NSCLC data. Panel A-D show the feature plot of the most differentially expressed gene in mast cells, fibroblasts, CD8 Tem and CD8 naive T cells in order. Panel E shows the cell type annotation result performed by CellTypist.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

3.4. Drug-target interaction and drug candidate repurposing

Here we investigate the feasibility for drug candidate repurposing based on the drug-target interaction. The workflow is shown in **Figure 5**. The potential targets were generated based on the differentially expressed genes from different cell types. The drug targets were generated from the potential drugs from the Comparative

Effectiveness of Drugs (CeDR), a comprehensive database featuring information on drugs that are undergoing clinical trials or have completed. 46 drug candidates were interacted with 32 proteins across 14 cell types. The top candidates are listed in **Table 1**, which shows the binding affinity between protein receptors and drug ligands in different cell types. A few drug candidates with the lowest binding score were shown the **Figure 6**. The top drug candidates include Dihydroergotamine, Glibenclamide, Astemizole and Isoflupredone. Dihydroergotamine, a drug primarily used for migraines, shows the lowest binding affinities across different cell types and targets. The lowest binding affinity is observed for Dihydroergotamine-CD3D interaction in T cells, a crucial component of the T cell receptor complex, suggesting potential immunomodulatory effects[24]. Glibenclamide, a drug used for type 2 diabetes, shows affinity for VWF and CA4 in endothelial cells. Another study suggested that Glibenclamide could have anti-thrombotic effects, possibly related to its interaction with VWF[25]. Astemizole, an antihistamine, shows affinity for EDNRB in endothelial cells and MRC1 in macrophages. Its role in these cells is not clear, but a study by Zhang et al. suggested that Astemizole could have anti-inflammatory effects, possibly related to its interaction with these targets[26]. Isoflupredone, a corticosteroid, shows affinity for CD8A in T cells. Its role in T cells is still not clear. These drug candidates show low binding affinities across different cell types and targets, suggesting potential therapeutic targets for NSCLC. However, further investigations are needed to fully understand their mechanisms of action and potential side effects.

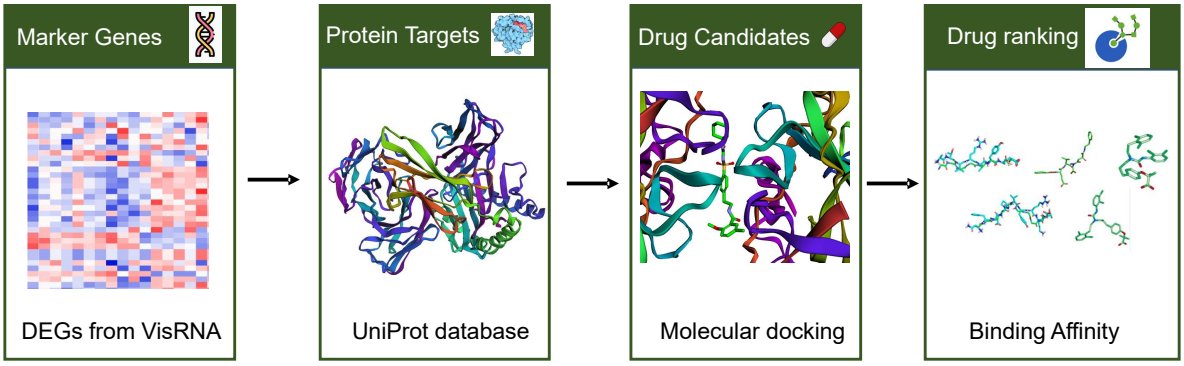

Figure 5. Workflow of drug candidate repurposing and ranking based on the differentially expressed genes from scRNA-seq NSCLC data.

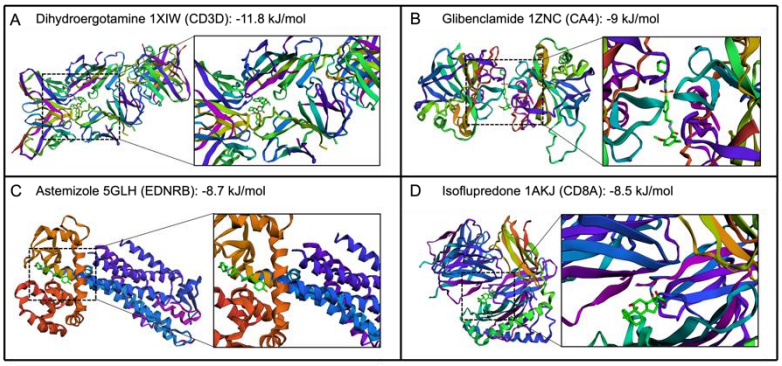

Figure 6. Molecular binding for drug and their drug targets with high affinities. Panel A shows the binding between Dihydroergotamine and 1XIW. Panel B shows the binding between Glibenclamide and 1ZNC. Panel C shows the binding between Astemizole and 5GLH. Panel D shows the binding between Isoflupredone and 1AKJ.

Table 1. Summary of top drug targets and drug candidates with their binding affinity score.

Celltype	Gene	Target	Drug	Affinity (kJ/mol)
Endothelial cell	3N3F	COL15A1	Dihydroergotamine	-8.2
Endothelial cell	5GLH	EDNRB	Dihydroergotamine	-11.3
Endothelial cell	1AO3	VWF	Dihydroergotamine	-10.9
Endothelial cell	1ZNC	CA4	Dihydroergotamine	-9.8
Endothelial cell	1AO3	VWF	Glibenclamide	-8.9
Endothelial cell	5GLH	EDNRB	Astemizole	-8.7
T cell	1XIW	CD3D	Dihydroergotamine	-11.8
T cell	1SY6	CD3G	Dihydroergotamine	-9.4
T cell	1AKJ	CD8A	Isoflupredone	-8.5
T cell	1AKJ	CD8A	Dihydroergotamine	-8.3
Macrophage	1EGG	MRC1	Astemizole	-8.1

Macrophage	1BHO	SPP1	Astemizole	-7.9
Fibroblast cell	1GQ5	PDGFRA	Dihydroergotamine	-9.2

3.5. Limitation and Future Works

While this study provides a valuable proof-of-concept for repurposing drug candidates through integrated analysis of scRNA-seq and molecular docking data, it is limited in scope. The current analysis focuses solely on NSCLC and uses a single dataset, restricting the broader applicability of the findings. Additionally, only a small set of compounds and targets were screened, likely omitting many potential interactions. Experimental validation through binding assays and in vivo studies is needed to confirm the predicted bindings. To build on this work, the analysis could expand to encompass diverse cancer types by incorporating additional datasets, enabling identification of therapeutic targets shared across tissues. Significantly expanding the drug and target screening could uncover more candidates for repurposing. Emerging nanotechnology may allow targeted delivery of top candidates based on identified molecular targets. The web platform could also be enhanced with more advanced and customizable visualization and analysis features. Finally, the approach could extend beyond cancer to other disease areas like neurodegeneration. This pioneering study establishes a framework for integrating multi-omics data to uncover therapeutic possibilities hidden within the heterogeneity of complex diseases. Significant expansion of the analysis and experimental follow-up are critical next steps for translating these computational findings into clinical treatments tailored to the intricacies of each patient's molecular profile.

4. Conclusion

VisRNA integrated the procedures from downloading scRNA-seq data, including dimensionality reduction, clustering, cell type annotation, and differential gene analysis, producing a user-friendly interface where all the data are transformed into visualizations for biological interpretations to find therapeutic targets. Machine

learning-based method cell type annotation is embedded in VisRNA. The differential gene analysis enables rapid identification of potential therapeutic targets. As a proof of principle, based on the DEGs from NSCLC, a few potential drugs such as Dihydroergotamine were proposed. In a sum, VisRNA is an interactive scRNA-seq data visualization platform for efficient data analysis and therapeutic targets discovery.

Acknowledgement

I want to express my sincere gratitude to my teacher for providing guidance and encouragement throughout this research project. Their belief in me gave me the confidence to take on an intimidating topic far outside my comfort zone. I'm also enormously appreciative of my family for their tremendous support on this journey. They listened patiently and helped me talk through complex concepts, even when the material was unfamiliar to them. Although tackling an advanced research paper presented many challenges, the experience has been incredibly rewarding. I'm thankful for everyone who accompanied me on this journey of academic growth. Their support gave me the motivation to push my abilities and explore new territories.

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).

[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).