Integrated Single-Cell Analysis and Visualization Software

Team 048: Sanya Badole, Hannah Bower, Rajan Ranjeet Tidke, Jacob Smith, Yashas Appaji, and Ryan Peng

---

1. Introduction

Single-cell RNA sequencing has revolutionized our understanding of cell sequencing in complex biological systems. It has allowed researchers to profile gene expression at unprecedented resolution. This technology has transformed fields ranging from developmental biology to precision medicine by revealing the diversity of cell types, states, and transitions that compose tissues. However, the analytical challenges associated with single-cell data have grown alongside dataset size and complexity. Our integrated platform addresses these challenges by unifying the workflow of single-cell analysis into a cohesive, scalable solution. By combining different computational methods with interactive visualization tools, our software enables researchers to preprocess, analyze, and explore large-scale single-cell datasets efficiently. This approach not only democratizes access to advanced analytical methods for non-programmers but also accelerates biomedical discoveries through improved visualization and integration capabilities.

2. Problem Definition

The current single-cell analysis landscape has some limitations that impede efficient and accurate data interpretation. Existing tools like Seurat and Scanpy are powerful, but suffer from three critical shortcomings. First, fragmentation of the analytical workflow forces researchers to navigate multiple tools for preprocessing, clustering, dimensionality reduction, and visualization, increasing complexity and reducing efficiency. Forcato et al, Conesa et al, and Han et al describe that this fragmented workflow increases complexity and reduces efficiency, while also emphasizing that there is no one-size-fit all approach [1][2][3][4]. Wu and Zhang and Xiang et al explained that techniques like t-SNE struggle with large datasets due to high computational costs [5][6]. Rutter et al 2019 pointed out that as datasets increase in size, these graphs struggle to accurately represent density [7]. While methods like SIMLR described by Wang et al. and net-SNE developed by Cho et al. show promise, they lack widespread adoption or integration into comprehensive platforms [8][9]. Static plots limit exploratory analysis. Narayan et al. demonstrated that tools like UMAP and t-SNE often fail to preserve local density information accurately, leading to misinterpretation of transcriptomic variability [10]. Additionally, as discussed by Amir et al. 2013, while visualization using viSNE is useful for non-conforming cells, its graphs have limited interactivity [11]. Limits in current practice include SPADE enables cellular hierarchy inference but lacks scalability for modern large-scale datasets [12] and Seurat v3 supports scRNA-seq integration but struggles with multimodal datasets [13]. These limitations highlight the urgent need for an integrated platform that addresses the technical complexity, scalability requirements, and visualization challenges inherent in modern single-cell analysis workflows.

3. Literature Survey

This project is relevant to Biologists researching cellular heterogeneity, Bioinformaticians with large datasets and clinicians looking for precision medicine. According to Peymani et al, Single Cell RNA-seq

can be used as a complementary diagnostic tool for genetic disorders, particularly when other methodologies like and whole-genome sequencing (WGS) fail to provide a molecular diagnosis [18].
The proposed platform will democratize single-cell data analysis by making advanced computational methods accessible to non-programmers. The biggest impact will be in the acceleration of biomedical studies. Forcato et al. demonstrated that integrated multimodal data analysis can accelerate discoveries in disease mechanisms [1]. Our platform will also significantly reduce time spent on preprocessing and data integration tasks. Lotfollahi et al. (2021) demonstrated that transfer learning approaches like scArches can perform reference mapping using four orders of magnitude fewer parameters than de novo integration while preserving biological variation [19]. We will also enable cross species knowledge transfer beyond Cho et al.'s work on cross-study knowledge transfer, recent benchmarking by Lotfollahi et al. (2021) showed that properly designed reference mapping can facilitate annotation transfer across species with approximately 84% accuracy [9] [19]. This enables researchers to leverage knowledge from well-characterized model organisms to study less-explored species. Our approach will also help with preserving biological variation, unlike conventional batch correction methods, which Tallulah et al describe as based on the assumption of shared biological conditions across cells [20]. Lotfollahi et al. (2021) demonstrated that transfer learning methods can retain disease variation when mapping COVID-19 samples to healthy references [19]. We will measure our success metrics by conducting user studies to assess usability, compare runtime and accuracy to current tools like Seurat or Scanpy. We will also have adoption metrics like downloads, citations in research publications, and community engagement.
Our biggest risks include technical complexity from integrating diverse algorithms into a unified interface as noted by many authors [1][3][4], as well as trouble convincing users of existing tools to transition from their more established tools. This is all balanced with excellent payoffs like a new transformative tool to simplify single cell analysis workflows, as well as commercialization opportunities as the software gains traction in academic and clinical settings.
We will need to account for computational infrastructure for pipeline development and testing, time for development, optimization, tool upkeep and documentation, example datasets for validation, and training materials development. Time will by far be the most expensive of them all, followed by computational resources. Duò and Robinson discuss the runtime and scalability of several single cell methods, which we will closely consider [21]. Out-of-pocket financial costs incurred will be **zero**.

4. Proposed Method

We have developed a working single-cell analysis pipeline capable of pre-processing data and visualizing it using various methods. The very first thing we did was look for a dataset to analyze. We decided on a Breast Cancer dataset that was publically available on GEO. The dataset consisted of over 100,000 individual single cells with their respective gene expression information. The data was in 10x Genomics format (barcodes, matrix, and features files), a common single cell data format, along with a metadata.csv file. The 10x Genomics data was converted to an anndata object based on standardized formatting practices and saved as an .h5ad file for uniformity. The library stacks we made the most extensive use of throughout this pipeline were - Scanpy, PHATE, TriMap, scvi-tools, Dash and D3.js. In the pre-processing stage, we performed Mitochondrial gene filtering (pct_counts_mt < 5), Low-quality cell filtering (n_genes_by_counts < 2500, min_genes ≥ 200) and Gene filtering (min_cells ≥ 3). After that we proceed with Normalization of the single cell data by setting target_sum=10000 using sc.pp.normalize_total(). We also do a Log-transformation using sc.pp.log1p().

Lastly, using sc.pp.highly_variable_genes(), we choose the top 2000 Highly variable genes (HVGs) which are our features for the next tasks, and scale this data with max_value=10 using sc.pp.scale(). This marks the end of the pre-processing stage, and we move on to feature representation and selection using Dimentonality reduction and clustering techniques.

We performed PCA for the top 2000 HVGs (with 40 components) which served as the base method for others. Then we implemented both 2-dimensional and 3-Dimensional UMAPs since it preserved neighbourhood structures. We also implemented t-SNE (with n_pcs=40, n_jobs=4), TriMap (on top 40 PCs), PHATE (n_components=2) which were useful for preserving global structures and trajectory inference. For clustering we employed Nearest neighbour graph with n_neighbors=10 and n_pcs=40 using sc.pp.neighbors(). Leiden clustering was also performed using sc.tl.leiden() .

The final processed dataset is then saved as **'processed_data.h5ad'.** All the intermediate representations (e.g., PCA, UMAP, t-SNE, PHATE, clusters) are saved in adata.obsm and adata.obs fields. The next phase of our workflow will use these components from the processed data and visualize them as an interactive dash web app. It is built with Dash and dash_bootstrap_components and has a modular layout using use_pages=True. It integrates D3.js and WebGL for interactive graph rendering. The launch command opens a local web server (default port 8050) with multiple tabs for each visualization. For this final report, only a basic scaffolding is complete, but this frontend is still critical for democratizing the platform—enabling biologists to interact with dimensionality plots and annotations without coding.

Our platform unifies the single-cell RNA-seq analysis pipeline—from quality control to interactive visualizations—within a Python-based ecosystem. Currently popular workflows, especially those in R (e.g., Seurat) are indeed powerful, but they lack flexibility in incorporating modern machine learning-based methods like variational autoencoders or density-preserving projections. Furthermore, our pipeline has been tested on multiple cell type datasets (both large and small) to ensure it is generalizable and scalable for future applications.The analysis and visualization scripts we have written will serve as the backend of our platform, which will serve to interactively visualize and summarize the results of several different datasets and types. This unified system is modular by design, allowing rapid testing of clustering algorithms, DR methods, or annotation strategies by swapping components with minimal effort.

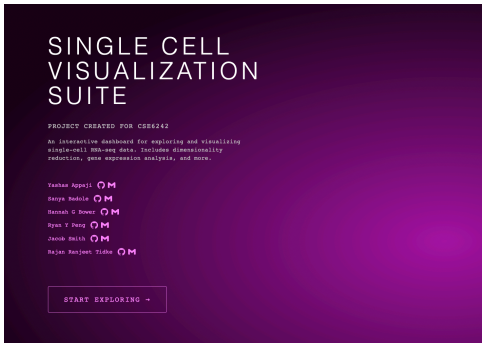


Fig. 1 - Homepage of the dash web app

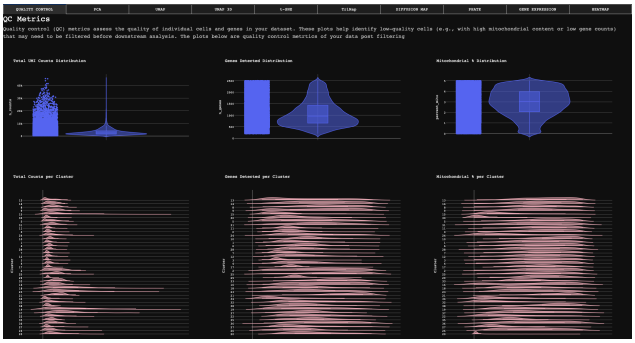
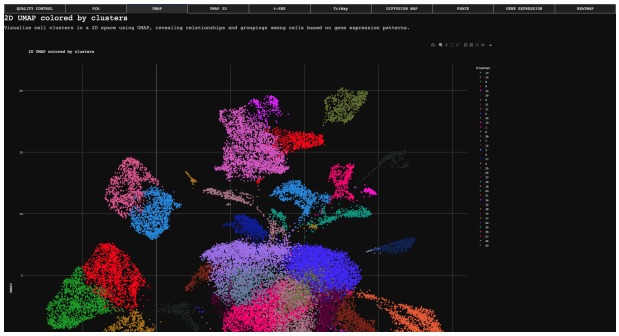
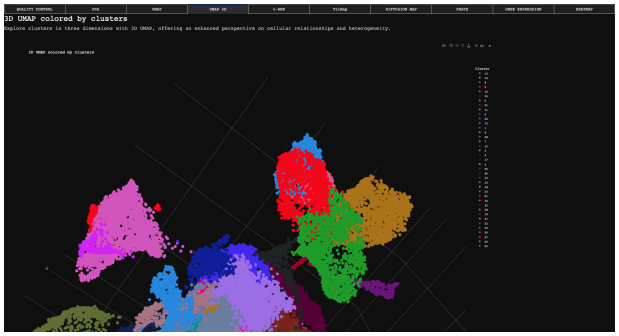


Fig. 2 - Plots of quality control metrics

mFig. 3 and 4 - Visualizing the cell populations using 2D and 3D UMAP

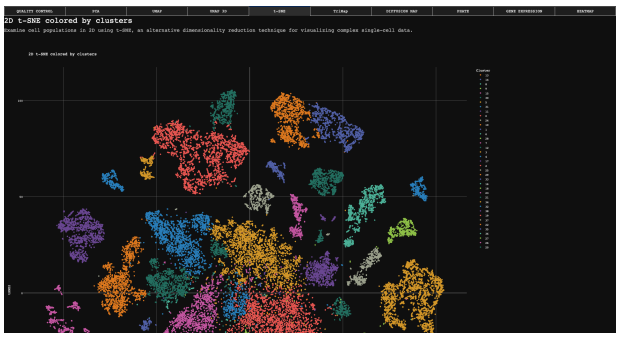



Fig. 5 and 6 - Visualizing the cell populations using 2D t-SNE and Trimap to preserve global structure

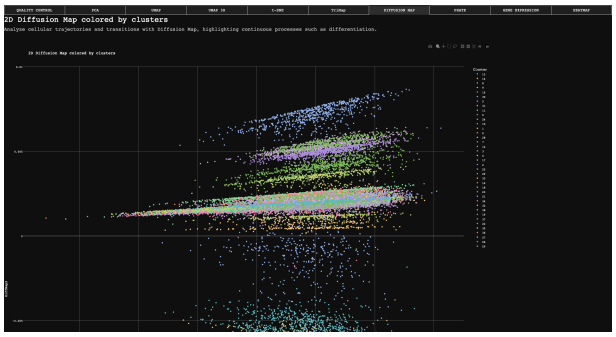
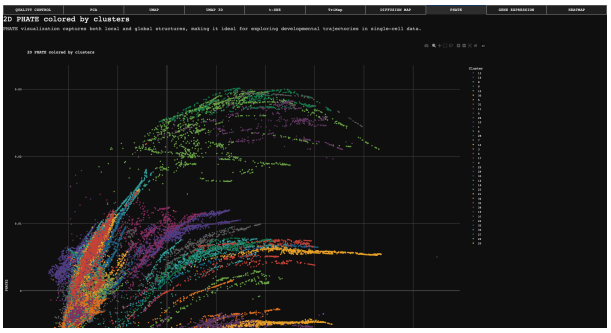



Fig. 7 and 8 - 2D diffusion map and PHATE visualization to explore cellular developmental trajectories

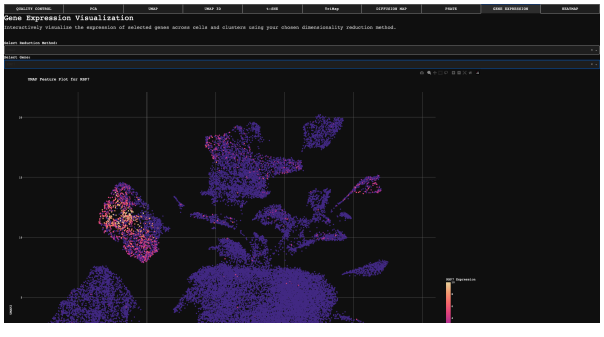
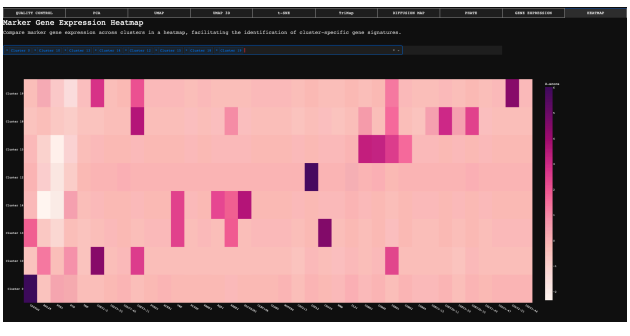



Fig. 9 - Visualizing the cell populations dimensionally reduced by a method (UMAP, t-SNE, TriMAP, Diffusion map pr PHATE) of choice and the gene of choice.

Fig. 10 - Visualizing the gene expression heatmap with a choice of selection of the cluster numbers.

5. Evaluation

Three ways we used to evaluate the success of our developed tool is runtime, testing multiple datasets and a user study. Runtime was about 20 minutes overall for the dataset we used, with the preprocessing segment taking about 19 minutes of the runtime and visualizations taking less than one minute. Testing across a range of devices with varying specifications showed consistent performance. We also tested our preprocessing scripts against multiple datasets to ensure the tool to be free of errors and not built too specifically to one dataset structure. Once we tested on about 15 datasets of varying datasets, we felt confident our analysis could handle a wide variety of .h5ad datasets.
Our user study was conducted on 10 students whose research focus is in single-cell RNA sequencing. They were asked to complete standard analysis and visualization tasks using our tool and provide feedback on usability, performance, and overall user experience. All participants were able to complete the tasks without needing to write code, and 90% reported that the interactive dashboard improved their understanding of clustering and trajectory results compared to static plots. None of the participants needed to add any additional code that was not specified in the .README file to run the analysis tool. Participants emphasized that the tool's low barrier to entry made it suitable for both new and experienced users. All noted that they were able to quickly understand how to operate the tool thanks to the intuitive layout and comprehensive README documentation.Students reported an average of 10/10 for usability, 9/10 for performance and a 10/10 for user experience. The interactive visualizations were frequently cited as one of the most valuable features. Users appreciated the ability to toggle between dimensionality reduction techniques (UMAP, t-SNE, PHATE, etc.) and explore gene expression dynamics in 2D and 3D with real-time interactivity. This feedback highlights our platform's potential to lower the barrier for non-programmers while maintaining analytical rigor.

6. Conclusions and Discussion

In essence, our project provides a streamlined and accessible platform designed to revolutionize single-cell RNA sequencing analysis. By integrating diverse computational methods and interactive visualization tools into a unified, Python-based ecosystem, our software empowers researchers—regardless of their programming expertise—to efficiently preprocess, analyze, and explore large-scale single-cell datasets. Our results from testing the software showed that the platform could pre-process datasets up to 100000 cells in just 20 minutes. Moreover, results from our testing, based on varied single cell datasets that could be pre-processed from the backend, and a user study indicates the usability of the dashboard for generating informative visualizations for inexperienced researchers. Our platform significantly democratizes access to sophisticated analytical techniques and accelerates biomedical discoveries through improved visualization and integration capabilities.

Despite its comprehensive capabilities, our platform currently faces several important limitations. The current version exclusively supports unimodal scRNA-seq data in the .h5ad format (AnnData objects), which restricts compatibility with alternative formats such as Seurat objects, Loom files, CSV matrices, legacy datasets stored in older formats, and direct imports from sequencing platforms using proprietary formats. Additionally, the platform lacks support for multi-modal single-cell data, including single-cell ATAC-seq for chromatin accessibility profiling, spatial transcriptomics technologies (e.g., Visium, Slide-seq, MERFISH), CITE-seq and other protein-measurement modalities, and multi-omic approaches that simultaneously measure multiple cellular properties. This limitation hinders integrated analysis of complementary data types that could provide more comprehensive biological insights. Furthermore, the

current version has limited integration with external resources, such as public reference datasets, cell atlases, automated cell type annotation tools, pathway enrichment frameworks, and external visualization or downstream analysis tools. During our user survey, some participants requested additional features, such as, customizable color schemes for clusters, and options to export interactive plots as high-resolution images or web embeds.

Our development strategy focuses on systematically addressing these limitations while expanding the platform's capabilities. First, we will implement comprehensive multi-modal integration, including a unified analysis framework for RNA-seq, ATAC-seq, and protein measurements, specialized analytical methods for each modality, cross-modality correlation techniques, spatial transcriptomics support with tissue context preservation, and tailored visualization tools. Second, we plan to incorporate sophisticated automated cell type annotation by integrating major cell atlas projects (e.g., Human Cell Atlas, Tabula Muris), deploying machine learning-based classification algorithms, enabling reference-based annotation with confidence scoring, and supporting custom annotation dictionaries for specialized cell types. Third, cloud-based infrastructure will be prioritized to provide scalable computing resources, secure data storage with access control, real-time collaborative environments, public data repository integration, and reproducible workflow sharing. Fourth, advanced visualization and reporting features will be developed, including publication-ready figure generation with extensive customization, interactive web-based result sharing, automated methodological reporting, and journal-specific formatting tools. Finally, we aim to extend the platform's utility in clinical research through compliance with data security standards, integration with electronic health records, specialized diagnostic workflows, biomarker discovery pipelines, and longitudinal sample analysis frameworks.

Our unified platform for scRNA-seq analysis represents a significant advancement in making sophisticated single-cell analytics accessible to the broader research community. By addressing key challenges of scalability, accessibility, and workflow fragmentation, we enable researchers to focus on biological questions rather than computational obstacles. While the current version has limitations in supported formats and modalities, our roadmap outlines a clear path toward a fully integrated single-cell analysis ecosystem. Planned additions—including multi-modal integration, automated annotation, cloud collaboration, and advanced reporting—will substantially enhance the platform's utility across academic and clinical settings. By evolving in response to user feedback and emerging technologies, the platform aims to accelerate discovery and translation in single-cell genomics.

All team members have contributed a similar amount of effort in this project.

References:

1. Forcato, M., Romano, O., & Bicciato, S. (2020). Computational methods for the integrative analysis of single-cell data. Briefings in Bioinformatics, 22. https://doi.org/10.1093/bib/bbaa042.
2. Pola-Sánchez, E., Hernández-Martínez, K. M., Pérez-Estrada, R., Sélem-Mójica, N., Simpson, J., Abraham-Juárez, M. J., Herrera-Estrella, A., & Villalobos-Escobedo, J. M. (2024). RNA-Seq data analysis: A practical guide for model and non-model organisms. *Current Protocols*, 4, e1054.https://doi.org/10.1002/cpz1.1054
3. Conesa, A., Madrigal, P., Tarazona, S. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016). https://doi.org/10.1186/s13059-016-0881-8
4. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinformatics and Biology Insights. 2015;9s1. doi:10.4137/BBI.S28991
5. Wu, Y., & Zhang, K. (2020). Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nature Reviews Nephrology, 16, 408-421. https://doi.org/10.1038/s41581-020-0262-0.
6. Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A comparison for dimensionality reduction methods of single-cell RNA-seq data. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.646936
7. Rutter, L., Moran Lauter, A. N., Graham, M. A., & Cook, D. (2019). Visualization methods for differential expression analysis. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2968-1
8. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2016). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature Methods, 14, 414-416. https://doi.org/10.1101/052225.
9. Cho, H., Berger, B., & Peng, J. (2018). Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks. Cell systems, 7(2), 185-191.e4. https://doi.org/10.1016/j.cels.2018.05.017.
10. Narayan, A., Berger, B., & Cho, H. (2020). Assessing Single-Cell Transcriptomic Variability through Density-Preserving Data Visualization. Nature biotechnology, 39, 765-774. https://doi.org/10.1038/s41587-020-00801-7.
11. Amir, E. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., & Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, *31*(6), 545–552. https://doi.org/10.1038/nbt.2594
12. Anchang, B., Hart, T., Bendall, S., Qiu, P., Bjornson, Z., Linderman, M., Nolan, G., & Plevritis, S. (2016). Visualization and cellular hierarchy inference of single-cell data using SPADE. Nature Protocols, 11, 1264-1279. https://doi.org/10.1038/nprot.2016.066.
13. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *cell*, *177*(7), 1888-1902.
14. Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P. Łabaj, P., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. Frontiers in Genetics, 14. https://doi.org/10.3389/fgene.2023.997383

15. Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W., Mohsen, H., Moon, K., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., Desai, A., Ravi, V., Kumar, P., Montgomery, R., Wolf, G., & Krishnaswamy, S. (2017). Exploring single-cell data with deep multitasking neural networks. Nature Methods, 16, 1139-1145. https://doi.org/10.1038/s41592-019-0576-7.
16. Manzoor, F., Tsurgeon, C. A., & Gupta, V. (2025). Exploring RNA-Seq Data Analysis Through Visualization Techniques and Tools: A Systematic Review of Opportunities and Limitations for Clinical Applications. *Bioengineering*, *12*(1), 56. https://doi.org/10.3390/bioengineering12010056
17. Katz, Yarden and Wang, Eric T. and Silterra, Jacob and Schwartz, Schraga and Wong, Bang and Thorvaldsdóttir, Helga and Robinson, James T. and Mesirov, Jill P. and Airoldi, Edoardo M. and Burge, Christopher B. (2015). Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics, 31(14),* 2400-2402. https://doi.org/10.1093/bioinformatics/btv034
18. Peymani, Fatemeh, et al. "RNA sequencing role and application in Clinical Diagnostic." *Pediatric Investigation*, vol. 6, no. 1, Mar. 2022, pp. 29–35, https://doi.org/10.1002/ped4.12314
19. Lotfollahi, M., Naghipourfar, M., Luecken, M.D. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* **40**, 121–130 (2022). https://doi.org/10.1038/s41587-021-01001-7