A comparison of single cell RNA sequencing visualization tools for multimodal timelapse analysis

CPSC 547: Project Proposal *October 21, 2022*

Kieran Maheden - maheden@student.ubc.ca
Brett Kiyota - brettckiyota@gmail.com

Introduction

An understanding of cellular processes is foundational to numerous fields in biology and the life sciences. Cellular processes are driven by the expression or activation of specific genes, encoded in the genome or DNA of a cell (Buccitelli et al., 2020). This DNA is "transcribed" into RNA, and in turn this RNA is then "translated" into protein. DNA and RNA are typically considered inert forms of information storage that allow for tight regulation of the functional gene product, the protein.

Due to how respective technologies have advanced over the past two decades, the most cost-efficient and data-rich way of characterizing global cellular gene expression is to measure the abundance of RNA molecules coming from a cellular population, commonly referred to as RNA sequencing (RNAseq). RNA sequencing provides count data for each gene identified (>50 000 genes typically) for each sample analyzed. Technical advances in the last 5 years have driven the advent of "single-cell RNAseq" (scRNAseq). With scRNAseq, count data is generated on a per-cell basis instead of averaging gene expression over thousands or millions of cells in a population. scRNAseq generates datasets that are significantly more complex to analyze and visualize, with the potential of >50 000 genes per cell with the number of cells ranging from a few thousand into the millions.

Further complicating the techniques is the introduction of "multimodal" experiments, where information at the DNA or protein level is also captured per cell (albeit at a much more information-sparse level). Incorporation of these different modalities into analysis and visualization tools adds an additional layer of complexity, with the standard visualization tools only publishing packages to achieve this in the last year (Hao et al., 2021). Due to the scale, high dimensionality, and complexity that multimodal datasets present, the task of extracting underlying biological meaning from such information involves navigating a multitude of challenging decisions at each step. Given that each choice can significantly influence downstream interpretations, and that each choice is highly context-dependent, visualization is a powerful tool that can help inform the decision-making process for all types of users.

Our group has some degree of familiarity with scRNAseq analysis using Seurat (Hao et al., 2021) and Scanpy (Wolf et al., 2018). BK commonly works with sequencing datasets for his primary thesis project. KM has analyzed his own datasets generated during the summer of 2022. Both group members have not obtained experience in integrating other modalities into RNAseq datasets.

Related Work

Recent years have seen an ever growing amount of scRNAseq papers include multimodal analysis, most clearly seen with multimodal scRNAseq being named Nature's method of the year in 2019. Further, one of the first examples of a multimodal technique,

CITEseq (Stoeckius et al., 2017) has received >1500 citations since its publication only a few years ago. As adoption of this technique grows, so too does the need for a suite of visualization tools capable of integrating the various modalities of information.

Integration of multimodal data into scRNAseq analysis has progressed alongside the technology required to obtain it. The two most common scRNAseq analysis and visualization suites, Scanpy for python and Seurat for R, have only recently released packages enabling this analysis (Hao et al., 2021, Bredikhin et al., 2022). Additional packages and tools have been developed outside of Scanpy and Seurat for multimodal analysis, however they have seen limited use due to their lack of integration with established pipelines (Forcato et al., 2021). To date, there have been no comprehensive comparisons between these tools - given the novelty of multimodal scRNAseq analysis and visualization, and proper comparison is imperative.

Data and Task Abstraction

Domain

Our project aims to conduct an analysis on the existing software tools for visualization that can be employed throughout the workflow of multimodal data analysis. If we were to consider scRNAseq data, the analysis pipeline typically involves the a pre-processing stage (e.g., quality control and feature selection, followed by the downstream analyses (cluster analysis, trajectory inference, expression heatmaps, etc…). Each step is typically accommodated by different visualization tools to guide the user towards the analysis choices that are most appropriate given the context of the dataset. The number of analysis steps can imaginally become quite convoluted once the other modalities are factored in. Thus, while the possibility of a generalized framework could be extremely attractive, such tools must overcome numerous challenges presented by large-scale and high-dimensional data.

Data

We will be exploring a Kaggle dataset (https://www.kaggle.com/competitions/open-problems-multimodal) that is focused on the developmental process of bone marrow stem cells as they differentiate into the various types of mature blood cells. The multimodal data is composed of information from approximately 300,000 cells and includes chromatin accessibility (DNA), transcriptomic profiles (RNA), and surface protein marker levels (protein). The information was collected at multiple time points, which adds a dynamic element that must be considered. In addition, the organizers of the Kaggle competition also inferred discrete labels for the specific type of each cell using a methodology established in a previous paper (Velten et al., 2017), which they provided as a supplemental resource to guide exploratory analysis.

In terms of explicit dataset size, there are approximately 20,000 cells upon stratification of the dataset based on time point. For chromatin accessibility, they were able to assess approximately 200,000 chromatin sites, where each value represents peak counts that have undergone the TF-IDF transformation (Term Frequency - Inverse Document Frequency). RNA gene expression levels are provided as a log1p-transformed (log transformation after adding 1 globally across values) count for each transcript, and the sequencing technologies capture

expression levels of approximately 20,000 genes. Lastly, the dsb (Denoised and Scaled by Background) normalized levels of 140 surface proteins are provided. Note that the respective transformations resulted in continuous data for each modality of information.

Task abstraction
While the primary goal of the Kaggle competition is geared towards developing machine learning models that can predict changes in RNA expression and protein levels as cells begin to differentiate into mature cells with a specific effector function, there are an abundance of other interesting biological questions that such a dataset can motivate. One particular area of interest revolves around the idea of assigning a discrete cell type label to individual cells (i.e., based on genetic information of a cell, we can assign a label to that cell based on where it may appropriately fit into existing knowledge of different cell types).

There are two major limitations with this current cell type labeling paradigm. The first limitation is that the practice of assigning a discrete label to a dynamic and continuous entity is inherently flawed, as cells are constantly changing in response to intrinsic and extrinsic factors. The second limitation is that cells are typically labeled solely based on transcript count information from scRNAseq data. However, the availability of information on the state of the cell's DNA (via chromatin accessibility) and protein levels presents the opportunity to devise a cell type labeling strategy that is more coherent with respect to the different layers of information, as well as how those different layers are subject to change over time. As such, we aim to evaluate the existing visualization tools that are commonly used to guide interpretations throughout this cell type annotation task. More specifically, our task will involve evaluating visualization tools that independently focus on each layer (DNA, RNA, and protein), as well as those that integrate information across the different levels. For the different visualization software tools, we can then analyze how the selected visualizations can influence domain-specific interpretations, which would allow us to better understand the nuances that must be considered with each approach.

Solution: your proposed infovis solution

Our project will primarily use R (ggplot2) and Python (matplotlib, Seaborn) to construct visualizations. Pre-existing software and toolkits will also be used for data processing tasks and further visualization; however, we will not be building any such software ourselves.

Use-case scenario
Given that the user interface is dependent on the explicit software used, our project will not be attached to any single user interface. However, possible visualization options include:

- Seurat (Hao et al., 2021)
- Scanpy (Wolf et al., 2018)
- WaddingtonOT (Schiebinger et al., 2019)
- Circos (Krzywinski et al., 2009)
- scMT-seq (Hu et al., 2016)
- MOFA+ (Argelaguet et al., 2020)

Kieran Maheden
Brett Kiyota

Milestones

Week	Kieran's Task(s)	Brett's Task(s):
8	• Download dataset and set up conda environment for Python (v3.9) and R (4.1)	
	• Related work research to identify software tools (focusing on scRNA-seq data) • Initial exploratory analysis to characterize dataset	• Related work research to identify software tools (focusing on multimodal data) • Data abstraction (how to integrate data across different layers)
9	• Finalize tasks; tasks abstraction • Determine list of software tools that will be used and download them all any dependencies	
10	• Seurat analysis of scRNAseq	• Scanpy analysis of scRNAseq
11	• Seurat analysis of multimodal	• Scanpy analysis of multimodal data
	• Discuss breakdown of additional software tools	
12	• Analysis of other software tools	• Analysis of other software tools
13	• (if necessary), revisit visualization tools and conduct further analyses	
14	• Formal writeup of paper • Slides for final presentation	

Bibliography

Argelaguet, R., Arnol, D., Bredikhin, D. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 21, 111 (2020). https://doi.org/10.1186/s13059-020-02015-1

Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol* 23, 42 (2022). https://doi.org/10.1186/s13059-021-02577-8

Buccitelli, C., Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* 21, 630–644 (2020). https://doi.org/10.1038/s41576-020-0258-4

Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. Brief Bioinform. 2021 Jan 18;22(1):20-29. doi: 10.1093/bib/bbaa042. PMID: 32363378; PMCID: PMC7820847

Hao Y, Hao S, Andersen-Nissen E, et al., Integrated analysis of multimodal single-cell data. Cell.

2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.

Hu, Y., Huang, K., An, Q. *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 17, 88 (2016). https://doi.org/10.1186/s13059-016-0950-z

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639-45. doi: 10.1101/gr.092759.109. Epub 2009 Jun 18. PMID: 19541911; PMCID: PMC2752132.

Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Gould J, Liu S, Lin S, Berube P, Lee L, Chen J, Brumbaugh J, Rigollet P, Hochedlinger K, Jaenisch R, Regev A, Lander ES "Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming." Cell. 2019 Feb 7;176(4):928-943.e22. Jan 31. https://doi.org/10.1016/j.cell.2019.01.006

Stoeckius, M., Hafemeister, C., Stephenson, W. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14, 865–868 (2017). https://doi.org/10.1038/nmeth.4380

Velten, L., Haas, S., Raffel, S. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* 19, 271–281 (2017). https://doi.org/10.1038/ncb3493

Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0