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

CPSC 547: Project Proposal *November 15, 2022*

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

Introduction

As the fundamental unit of life, cells are dynamic biological entities in which molecular and chemical processes are occurring in a constant state of flux. The amount of information packed into every single cell in the human body is incomprehensible. With approximately 3.2 billion nucleotides (i.e., DNA bases – {A, C, G, T}) in the genome, a cell's DNA is estimated to encode for somewhere between 20,000 and 25,000 genes. Each gene can be transcribed into an intermediate state made up of ribonucleotides (RNA), which can in turn be translated into proteins (Buccitelli et al., 2020). While DNA and RNA are typically considered inert forms of information, protein molecules are largely responsible for governing the functionality of a cell, and recent estimates suggest that a cell can contain as many as 42 million proteins at a given point in time.

This flow of information, from DNA to RNA to protein, describes what is referred to as the central dogma of biology, and the many levels and mechanisms of regulation that can occur along this flow result in a magnitude of complexity that the human brain cannot begin to grasp. Despite such challenges, there have been countless large-scale efforts to characterize the different levels (or modalities) of information within each cell. This is because having a deeper understanding of complex biological systems at their most basic unit (i.e., the cell) can facilitate novel insights into cellular processes that can potentially lead to innovations in medical and biological sciences.

Juxtaposed by the abundance of information that is present in each cell is their microscopic nature, which presents numerous challenges in quantitatively measuring different aspects of a cell. However, with the rapid advancement of sequencing technologies over the past two decades, large-scale and affordable DNA and RNA sequencing experiments have become widely accessible. As such, sequencing of a cell's genetic information has emerged as one of the most powerful methods for studying cell biology. In particular, cells are often quantified in terms of their gene expression profiles, where "expression" refers to the process by which a gene is transcribed from genomic DNA into an RNA molecule (which can be later translated into a protein). By measuring the abundance of RNA molecules of all of the genes within each cell (referred to as single-cell RNA sequencing, or scRNA-seq), we effectively gain a "snapshot" of the state of each cell in terms of their respective gene expression profiles, which can allow inferences to be made about the biological system of interest. One common use case of scRNA-seq information is to use each cell's gene expression profile as a basis for determining the type of the cell, where the rationale for such a labelling task is that differences in cell effector function should be reflected by corresponding differences in gene expression.

Recent advancements have extended scRNA-seq technologies to also permit the simultaneous quantification of a cell's gene expression with either DNA or protein information. Termed "multimodal" sequencing experiments, the ability to quantify a cell in terms of multiple layers of information (DNA:RNA or RNA:protein) offers a lot of promise from the standpoint of understanding how the different layers of information work together to shape a cell's overall function. However, this also presents many challenges from an analytical standpoint, as the complexity and number of attributes can increase substantially. Note that to our knowledge, there are no current technologies that are able to simultaneously and efficiently capture all three layers of information (DNA, RNA and protein) at single-cell resolution.

While visualization has been an extremely powerful tool in guiding the analysis and interpretation of gene expression information from scRNA-seq experiments, its application in the context of multimodal datasets is in its infancy (Hao et al., 2021). As such, our project aims to conduct an analysis on the visualisation softwares that can be applied to these multimodal datasets. In particular, we will analyze the effectiveness of existing software tools for the task of cell type labelling on a multimodal dataset that consists of RNA:protein information modalities at multiple time points.

Our group has some degree of familiarity with scRNA-seq analysis using field-standard visualization tools such as 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.

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 set 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.

At the core of most of these tools is the utilisation of non-linear dimensionality reduction (DR) techniques, most commonly t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (van der Maaten and Hinton, 2008; McInnes et al., 2018). Using DR, single cell experiments are often represented as a 2D scatterplot, with each cell's data (count data on up to as many as 20000 genes) as a point,

with more similar cells clustering together as a result of DR. After building these DR plots, the other modalities (non-RNA data) are often overlaid, providing additional insight.

DR via t-SNE and UMAP has seen significant usage outside of scRNAseq data analysis. The clearest demonstration of the versatility of these tools is seen in their original publications and their application to face image datasets (Olivetti faces), object datasets (COIL-20), handwriting image datasets (MNIST), word/phrase datasets (Google News Word Vectors), and a variety of biological datasets (flow cytometry and scRNAseq).

Data and Task Abstraction

Data
We will be exploring a Kaggle dataset that follows the developmental process of bone marrow stem cells (mobilized peripheral CD34+ hematopoietic stem and progenitor cells) as they differentiate into various types of mature blood cells (Velten et al., 2017). Cells were sampled from 4 healthy human donors (Day 1), and then measurements from these sampled cells were taken at 5 time points over a 10-day period (Days 2, 3, 4, 7 and 10). The dataset is comprised of information from two multimodal sequencing technologies:

10x Genomics Single Cell Gene Expression with Feature Barcoding technology (CITEseq): measures gene expression (RNA) and surface protein levels for each cell. RNA gene expression levels representing the abundance of RNA molecules for each gene are provided as the first modality of information for each cell, where the data have undergone a global (library-size) normalization and log1p transformation. This information is paired with cell surface protein levels that have undergone a denoised and scaled by background (dsb) normalization. Note that there are only 140 surface level proteins that are captured with this CITEseq technology, but the included proteins are known to be involved in important developmental processes in bone marrow stem cells.

Also provided in the dataset are cell type labels that were inferred based on RNA expression information using a method established in a previous paper. Each cell is labelled as one of the following cell types:

- Mast Cell Progenitor (MasP)
- Megakaryocyte Progenitor (MkP)
- Neutrophil Progenitor (NeuP)
- Monocyte Progenitor (MoP)
- Erythrocyte Progenitor EryP)
- Hematopoietic Stem Cell (HSC)
- B-Cell Progenitor (BP)

Data Abstraction
The RNA expression data is provided as a flat table with 70,988 items and 22,050 attributes. Each cell acts as a key to a particular item (row in the table), with a number of value attributes that represent different genes. Notably, the value for each item-attribute pair is a quantitative value representing the normalized RNA expression counts. Thus, each attribute encodes quantitative information.

In addition, there is a metadata table, which includes 281,528 items and 5 attributes. The attribute (with its type) can be stratified as follows:

- cell_id (categorical type): a unique identifying alphanumeric string that is assigned to each cell in the dataset.
- day (sequentially ordered quantitative type): represents the time point at which sequencing measurements were taken.
- donor (categorical type): a unique identifying number that is assigned to the 4 healthy adult donors.
- cell_type (categorical type): the inferred cell type label for each cell.
- technology (categorical type): the sequencing technology used (note that our project is focused on the CITEseq technology).

Such metadata can be mapped to the multimodal information by the cell_id attribute, which permits stratification of the data based on the remaining 4 attributes. Consequently, the dataset also consists of a time-varying semantic. This is shown in the Figure 1 bar chart that depicts the number of cells included in the dataset when categorized by cell type and day.

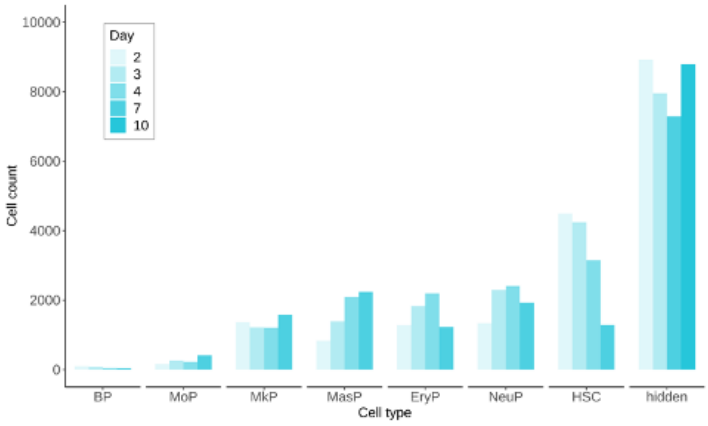

Figure 1: cell counts by cell type and day.

Task

Our project aims to evaluate the task of assigning a label to each cell. Cell type annotation is a highly disputed topic, due to the inherent limitations of assigning a discrete label onto continuous data that is derived from a dynamic biological entity. Conventionally, cell type labelling is performed by considering only a small subset of genes that are “known markers” for pre-defined cell types (i.e., certain cell types can be characterized by the expression of specific genes, and those specific genes can be denoted as “markers”), and labels are assigned to cells based on similarities in gene expression relative to those known marker genes. However, such a labelling paradigm has major flaws in that it does not consider all available gene expression information (additionally, it typically only considers a single modality of information, in RNA), it does not account for temporal change in gene expression patterns, and it’s restricted to labelling cells based on known cell types (i.e., you can never identify new cell types).

This dataset presents an opportunity to not only evaluate existing cell type labelling strategies, but also to investigate strategies that consider multiple modalities of information

(RNA and protein) and how they can vary over time. In particular, we will analyze how existing software tools can influence the interpretations of cell type labels, and whether those interpretations are subject to change depending on the type of information (RNA, protein, or both) and the time at which the measurement was taken.

Task Abstraction
With respect to the three levels of actions that tasks can be abstracted to, the highest-level action (Analyze) of cell type labelling would be to consume the cell-specific information in the multimodal dataset, as it is not well-understood. In doing so, we aim to discover new insights that can allow us to formally evaluate the effectiveness of existing software tools for our task of cell type labelling. The mid-level goal (search) will entail exploring characteristics of individual cells with no prior notions of their respective locations. These locations may manifest as relative outliers in a scatterplot with respect to the rest of the cell population or patterns in time-series variation plots. At the lowest-level user goal (query), we will attempt to summarize this cell type labelling task with respect to every cell in the population. It is necessary to consider all cells in a population because the cells can only be compared relative to other cells in the population, which is particularly important when identifying potential rare cell types (i.e., if a very small subset of cells are distant from the rest of the population).

In terms of the four kinds of abstract targets, our task will involve looking for trends amongst all of the data, as well as potentially topological patterns in network data. More specifically, we hypothesize that certain visualization idioms may delineate populations of cells with higher resolution than others when evaluating trends such as clusters. For example, while the provided cell type labels include only 8 cell types, we expect some methods may identify many more cell types (e.g., 30 different labels). With respect to network data, understanding the topology of protein interaction networks could also provide insights into possible sub-populations of cells.

As for the actual design of vis idioms, the high number of attributes will likely necessitate the utilization of multiple families of design idioms. Reducing, through filtering, aggregation and dimensionality reduction will all be employed throughout the analysis project as those form the basis of many existing software tools. In addition, juxtaposition and superimposition in the faceting family, and color/size/shape from the map family will also be incorporated into our analysis of visualization idioms.

Methods and Tools

This dataset has not been previously analyzed by BK or KM. For our project, one of the major themes for the visual analysis idioms that were selected was the ability to handle large-scale, complex datasets. Given that the continual improvement in sequencing technologies will facilitate increasingly large-scale multimodal experiments, there is a growing need to analyze the existing visualization software tools in the context of large-scale datasets. The reason why manipulation and other interactive visual data analysis idioms will not be considered for this project is that the lack of adoption of interactivity in the scientific community has resulted in relatively few software tools for the interactive analysis of multimodal sequencing datasets.

One of the major visual idioms will be scatterplots matrices, particularly ones that visualize data that has undergone some form of visual data analysis idiom such as dimensionality reduction (PCA and/or UMAP). With only 5 time points and 8 cell types to consider, these attributes are few enough in number to allow for the effective juxtaposition as rows and columns in a matrix, respectively. Moreover, such dimensionality reduction methods will be able to capture all of the data with respect to each time-cell-type pair.

Another visual data analysis idiom that will be heavily used is aggregation. With approximately 20,000 cells at each time point, an effective approach for reduction may be to aggregate information across cells with respect to cell type labels. This can simplify the comparisons across different cell types by reducing the density of information in the plots.

Color and superimposition will also be utilized to highlight specific trends that cell types may have with respect to the entire population of cells. In certain cases, ordering and filtering may also be applied to focus on specific subsets of genes (e.g., only consider genes with the highest RNA expression level in a particular analysis).

The primary tools that will be used for our analysis are Seurat and Scanpy. These are two of the most popular frameworks for the analysis of single-cell sequencing information, making them clear choices for our analysis on existing visualization tools. While Seurat/Scanpy will be used in tandem to produce a majority of the visualizations, we will also be evaluating “stand-alone” softwares for any remaining visual idioms that are not supported by either Seurat/Scanpy.

Cytoscape is on such software that will be utilized to construct protein network visual idioms, as it has been shown to be an extremely powerful tool that can produce large-scale network visualizations of protein and gene expression data.

Note that further stand-alone tools will be determined in the Milestones leading up to the final presentation.

Analysis

Initially, we sought to characterize the entire population of cells in terms of their gene expression profiles when juxtaposed by day (rows) and cell type label (column).

Figure 2 depicts a scatterplot matrix where the quantitative gene expression tables have undergone dimensionality reduction with PCA, followed by manifold learning with UMAP. The color channel is also used alongside superimposition to highlight the respective cell types against the entire population.

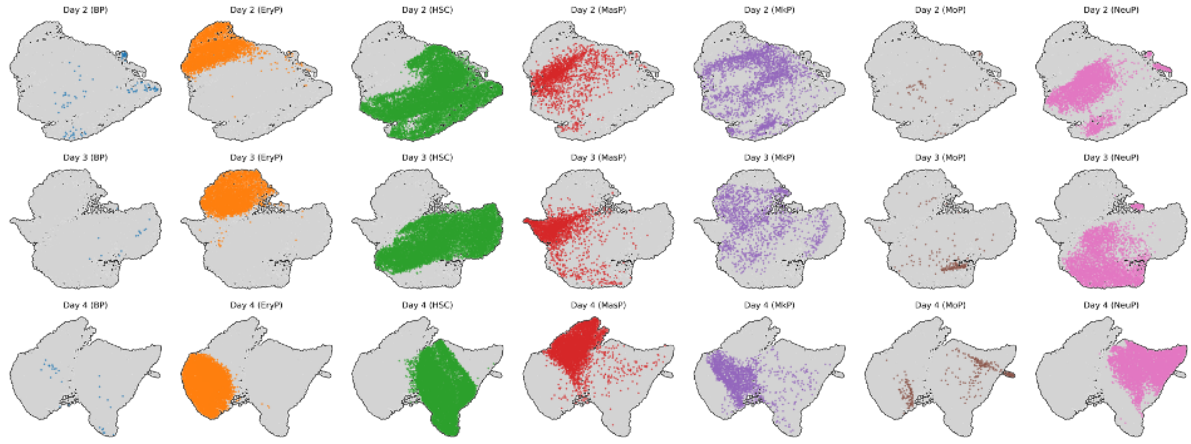

Figure 2. UMAP scatterplot matrix of gene expression profiles juxtaposed by day (rows) and cell type (columns)

Figure 3 depicts a similar UMAP scatterplot matrix for the surface level protein levels of each cell, which shows similar overall trends when compared to the gene expression profile.

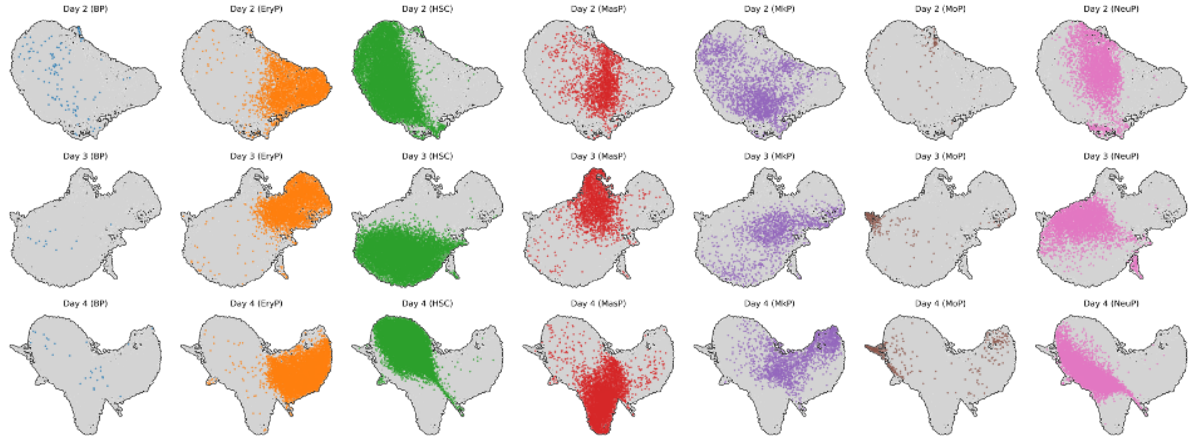

Figure 3. UMAP scatterplot matrix of surface protein levels juxtaposed by day (row) and cell type (columns)

From the above UMAP scatterplot matrices (Figure 2 and 3), we can observe the entire population of cells for each day-cell type pairing. In particular, we can see how at day 2, there is little separation across the different cell types, which suggests that their gene expression profile and protein levels are quite similar. However, for the measurements taken at day 3 and 4, we can see a modest degree of separation between cell types progressively emerge.

The B-cell progenitor (BP) cells were quite scarce and scattered throughout the population of cells, especially in terms of surface protein levels. Overall, the other cell types appear to have a more clear delineation toward the latter two measurements. One observation that the superimposition data analysis idiom enabled was that the cell populations for each cell type appear more scattered in the UMAPs that were based on surface protein levels. This may reflect the tendency to label cells solely based on gene expression information, rather than incorporating additional modalities of information such as protein levels.

Following the global analysis of all cells, a heatmap visual encoding of ranked gene expression was provided in Figure 4. This visual idiom involved ordering gene expression levels for each cell type, and filtering to include only the top 10 most differentially expression genes (i.e., positive or negative).

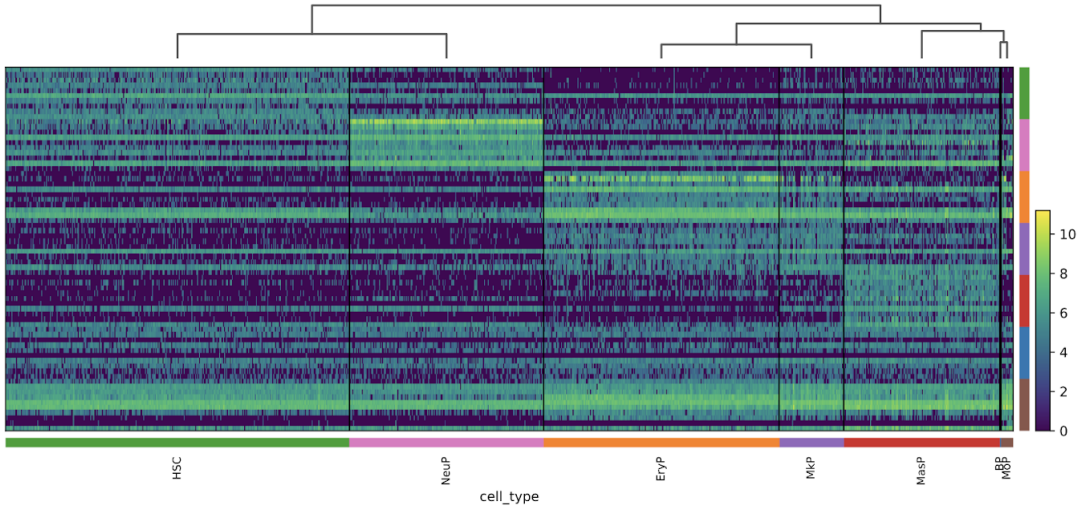

Figure 4. Heat map of ranked gene expression by cell type.

The resulting heat map depicts the fold change of gene expression levels of the top 10 differentially expressed genes in each cell type. Moreover, a dendrogram is provided to model the hierarchical relationship that each cell type has based on the selected genetic profiles. There is a clear trend in expression levels that can be observed with respect to each cell. While encoding less information than the UMAP scatterplot matrices, the above heatmap still depicts a dense amount of information for certain cells. However, this comes at the expense of losing consideration for the expression levels for genes not shown.

To get increasingly specific, the dotplot below (Figure 5) depicts a similar ranking of gene expression levels as the heat map above; however, by varying the size of dots, there is a more clear distinction between the relative magnitude of expression levels within each cell type. These latter two plots also suffer from the inability to capture multiple time points (without explicitly plotting each time point independently). This reflects a general lack of support for time-series sequencing datasets.

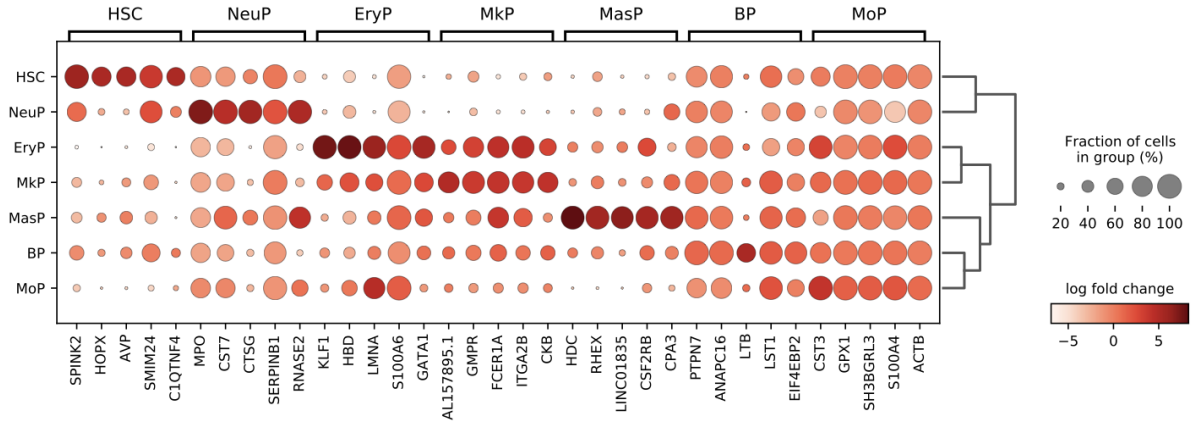

Figure 5. Dotplot of ranked gene expression profiles by cell type

One of the main reasons this dataset was selected was due to the fact that there were multiple modalities of information capture on a per-cell basis. Thus, in order to characterize a cell in terms of both its gene expression profile as well as its surface protein levels, an initial correlation scatterplot matrix was constructed (Figure 6). In particular, for the 34 RNA:protein pairs that were identified, there is a modest correlation between the gene expression level and protein level. This correlation did not appear to change depending on the time point at which the measurements took place nor based on the cell type.

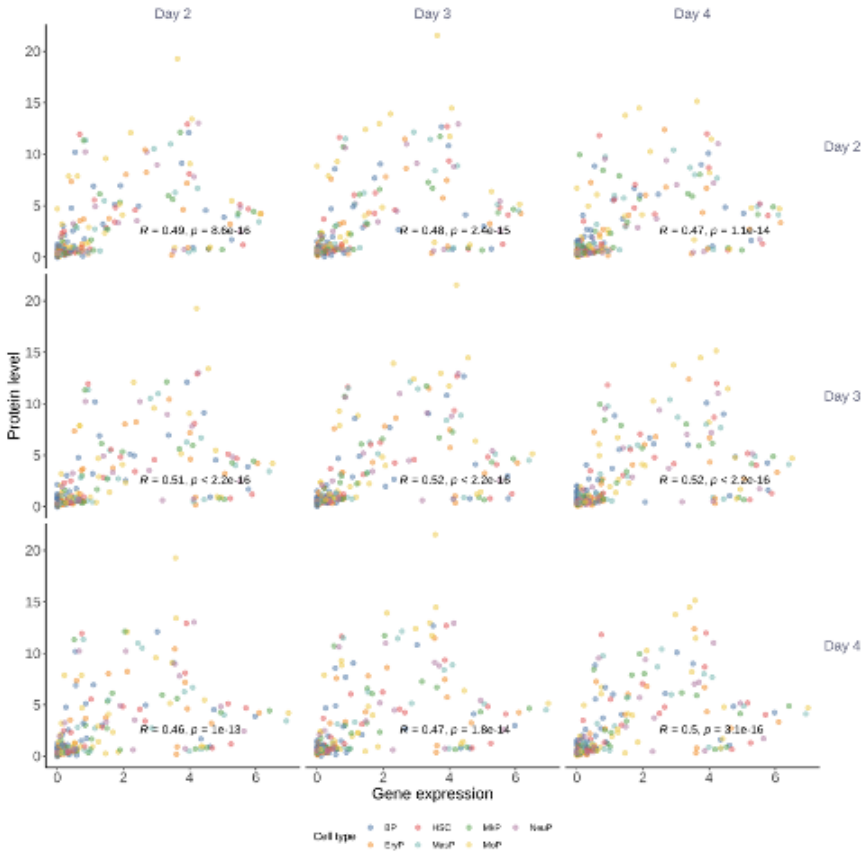

Figure 6. Evaluating the correlation between gene expression levels and surface protein levels for 34 RNA:protein pairs.

Milestones

Task	Est. Completion Date	Actual Completion Date	Est. Hours	Actual Hours
Download dataset and establish Conda environment for Python (v3.9)	2022/10/26	2022/10/26	2 hours	2 hours

and R (4.1)				
Related work research to identify software tools focusing on multimodal scRNA-seq data	2022/10/26	2022/10/26	4 hours	5 hours
Determine list of software tools that will be used and download them all any dependencies	2022/11/02	2022/11/02	2 hours	2 hours
Seurat and ScanPy analysis of dataset	2022/11/09	2022/11/13	4 hours	5 hours
Explore stand-alone multimodal exploration tools	2022/11/23		5 hours	
Perform analysis via stand-alone tools	2022/11/30		6 hours	
Iterate on ScanPy and Seurat analysis, comparing to stand-alone tools	2022/12/07		2 hours	
Formal writeup and presentation slide creation	2022/12/14		5 hours	

Discussion and Future Work

Bibliography

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.

McInnes L., Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv, (2018). https://doi.org/10.48550/arXiv.1802.03426

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

Van der Maaten L, and Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 9, 2579-2605 (2008).

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