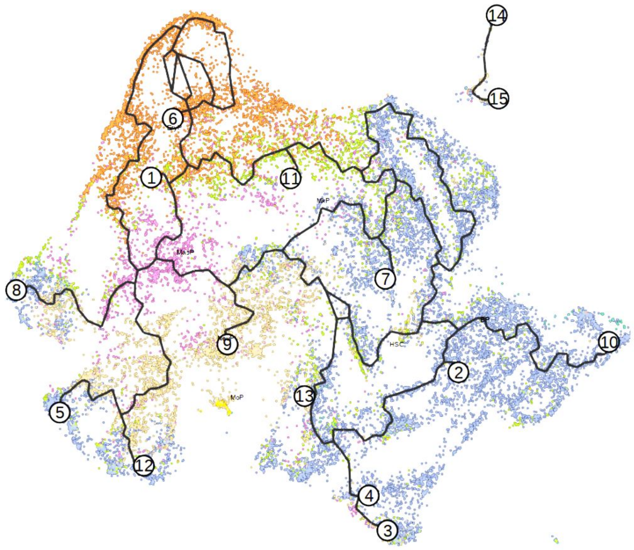
Visualizing single cell transcriptomes from timecourse data

Brett Kiyota, Kieran Maheden



Abstract—The ability to blindly characterize a biological system with single-cell RNA sequencing has revolutionized cell biology. By capturing RNA molecules and generating count or abundance data per cell, biologists are capable of generating massive datasets, informing them of cellular states and processes. One challenge now facing the field of single-cell sequencing is how best to analyse and visualize the abundance of data. Several tools have risen to prominence in the field, but no direct comparison or investigation has been performed on their ability to perform in the context of visualization. In this body of work, we explore a recently released dataset of bone marrow stem cells during differentiation, and examine common analysis and visualization tools (Scanpy, Seurat, Monocle3, and PAGA) in their ability to communicate meaningful biological insights. Ultimately, we highlight gaps in existing visualization idioms used by these tools through our own analysis and that of others, and propose our own potential solution.

Index Terms—scRNA-seq, Dimensionality Reduction, Transcriptomic time-course analysis, Pseudotime trajectory inference

1 Introduction and Domain Background

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 [3]. 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 [6].

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 (often referred to as the genome for DNA, transcriptome for RNA, or proteome for protein). 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 in experimental costs have enabled researchers to greatly expand the number of cells included in a given scRNA-seq experiment. One approach to leverage this is to perform time-course analysis, where cells are collected at regular intervals during a biological process such as differentiation, the process where stem

- *Brett Kiyota is a MASc student in the lab of Nozomu Yachie - brettckiyota@gmail.com.*
- *Kieran Maheden is a PhD Candidate in the lab of Nika Shakiba - maheden@student.ubc.ca.*

cells become more mature cells. In such experiments, analysis becomes focused on understanding the dynamic changes that occur from time point to time point.

Visualization of scRNA-seq experiments is central to single-cell analysis and exists as a powerful tool in guiding the interpretation of gene expression information from scRNA-seq experiments. Given the vast amount of data generated from such work, researchers often take an exploratory approach to characterizing cellular states. Initial exploration and visualization is typically performed using one of two common packages - Seurat or Scanpy [5, 21]. Both of these packages contain a variety of tools for plotting and exploring single-cell data, and have become widely adopted by the community. Given the increasing commonality of scRNA-seq with and without time course data and the relative immaturity of the field, we set out to understand the strengths and limitations of existing and widely adopted visualization tools.

Our own familiarity with scRNA-seq analysis is undeveloped. KM has a modest amount of experience with running Seurat packages on data that was generated in-house during his PhD work to date. BK has some exposure to analysing sequencing datasets, but has not explored scRNA-seq packages himself. Prior to this project, KM had not examined the analysed dataset. BK had initially identified the dataset by hearing about the associated Kaggle competition, but had not performed any analysis.

2 Data and Task Abstraction

2.1 Data

This analysis project explored a scRNA-seq dataset that was published on Kaggle, an online community for data scientists and machine learning enthusiasts focused on exchanging ideas and solving data science challenges. The dataset was part of an open challenge in 2022 to predict how DNA, RNA and protein measurements vary in single cells. For our analysis, we extracted gene expression and protein measurements from the overall dataset, which 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. Cells were sampled from 4 healthy human donors at day 1, and then these cells were allowed to grow and differentiate in a media for 3 days. A subsample of cells were collected each day (day 2, 3, and 4), which were subsequently characterized by a sequencing technology to obtain a profile of their gene expression and protein levels.

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) [16] was the technology used to quantify the molecular readouts of individual cells. For the 70,988 cells that were collected across the 3-day period, information was gathered for 22,050 genes and 140 proteins. Gene expression levels reflect the abundance of RNA copies for each gene that have undergone a global normalization and log1p transformation. Note that the log1p transformation is the natural logarithm of one plus the RNA count information, where the plus one ensures that taking the log-transform does not result in an undefined value. For each cell, this information is paired with cell surface protein levels that have been subjected to a denoised and scaled by background (dsb) normalization [12]. The dsb normalization is performed to treat technical noise that is introduced by the droplet-based sequencing technology [9] upon which CITE-seq is based.

The organizers of the Kaggle competition also provided a cell type label for every cell, which was inferred with RNA gene expression values using a method established in a previous paper [20]. In particular, each cell was labelled as one of the following mature blood 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)

Any cell that could not be categorized as one of the above cell types was discarded from analysis. Note that a progenitor cell is a term used to define a cell that has not fully differentiated into its mature cell type; however, it is far enough into its differentiation trajectory such that it has ”committed” to maturing into that target cell type. In contrast, HSCs are in a multipotent state, meaning that they are still capable of developing into all types of blood cells.

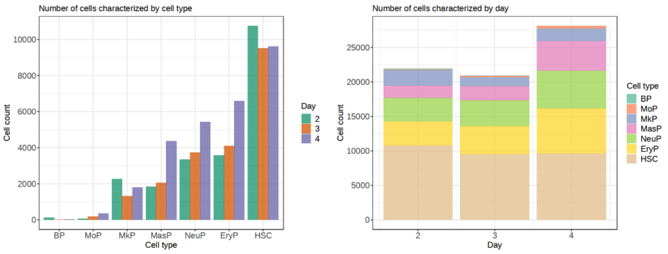


Fig. 1. Number of cells collected by cell type and day. **Left**: a grouped bar chart that depicts the breakdown of the number of cells collected across the three time points for each cell type. The cell types were ordered by the total number of cells collected in the 3-day period. **Right**: a stacked bar chart that shows the total number of cells collected at each day, with the color channel used to stratify those total counts by cell type.

Fig. 1 depicts bar plots that provide a breakdown of the number of cells that are collected across the 3-day period by cell type and day. Notably, there were very few BP and MoP cells collected and subsequently identified in this experiment. At each time point, a majority of cells were categorized as HSCs, which suggests that a large portion of the cell population have yet to commit to one mature blood cell type lineage. The left panel in Fig. 1 does indicate a relative decrease in HSCs after day 2, which is reflected by an increase in MoP, MasP, NeuP, and EryP cells at day 3 and 4. Over 20,000 cells were collected at each time point (Fig. 1).

2.2 Data Abstraction

Table 1. Metadata for molecular readout information.

Attribute	Type	Range
cell_id	categorical	NA
day	ordered quantitative	[2,4]
donor	categorical	N/A
cell_type	categorical	{MasP,MkP,NeuP, MoP,EryP,HSC,BP}
technology	categorical	{multiome,citeseq}

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 all of the genes. Notably, the value for each item-attribute pair is a quantitative value representing the derived expression values following scaling, normalization and transformation. Thus, each attribute encodes quantitative information.

The surface protein level data is also provided as a flat table with 70,988 items and 140 attributes. In this case, each cell is an item whose attributes correspond to the protein levels. The item-attribute pairs are quantitative values.

In addition, there is a metadata table, which includes 70,988 items and 5 attributes. The attribute (with its type) information can be found in Table 1.

- 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 CITE-seq 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.

2.3 Task

While highly convoluted, biological systems are extremely well-organized at a cellular level, and the clear structure and specialization within these cell populations is largely what enables organs and tissues to respond to stimuli in a coordinated manner. While the function of a given cell may be obvious when viewed within its natural environment, practically every sequencing workflow involves processing steps that result in the loss of this biological context. Consequently, how to derive meaningful biological insights from the molecular readouts obtained via single-cell sequencing technologies represents the essence of single-cell analysis.

The notion of assigning a biologically meaningful label to each cell may seem like a straightforward task; however, the dynamic nature of cells presents many challenges with this practice due to the inherent limitations of attempting to assign a discrete label onto continuous data. Despite the fact that cell type labelling is a highly disputed topic, it remains a key fixture in many analysis pipelines since this categorization step enables comparisons between different groups, or clusters, of cells.

As such, our project aims to evaluate the task of assigning a label to each cell. Conventionally, cell type labelling is performed by considering only a small subset of genes that are known “markers” for previously characterized cell types. For example, cells in our heart can generally be characterized by the expression of a specific subset of genes, and those specific genes can be leveraged to discern heart cells from other types of cells. Such labels are generally assigned to cells based on gene expression information. However, such a labelling paradigm has a number of flaws: (1) this paradigm does not consider all available gene expression information, (2) it only considers a single level of information, (3) it does not account for temporal change in gene expression patterns, (4) and it’s restricted to labelling cells based on known cell types (i.e., you can never identify new cell types).

The process of cell differentiation of blood cells is an extremely well-studied topic because of its importance in every organ and organ system in the body. As a result, this data set presents a great 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.

2.4 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 scatter plot 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 their molecular profile).

In terms of the four kinds of abstract targets, our task will involve looking for trends amongst all of the 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 7 cell types, we expect some methods may identify many more cell types (e.g., 30 different labels). In treating the assigned cell types as labels on which to compare clusters of cells, we expected to observe the qualitative separation of clusters to evolve over the 3-day period.

As for the actual design of visualization 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. Another common design idiom will be derivation, as there are many processing steps that involve subjecting a cell’s molecular information to various statistical tests, which are subsequently visualized to identify interesting patterns. 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.

3 Related Work

Recent years have seen an ever growing amount of scRNA-seq papers include more complex experimental design and analysis while still heavily leaning on basic tools such as Seurat or Scanpy. At the core of most of these tools is the utilisation of non-linear dimensional reduction (DR) techniques, most commonly t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) [8, 11]. Using DR, single-cell experiments almost always are represented as a 2D scatter plot, 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 or metadata (non-RNA data, time-course data, trajectory analysis) are often overlaid, providing additional insight.

Given that our dataset comes from the blood compartment of the body - one of the most well-studied stem cell systems - we also explored other groups’ attempts to analyze and visualize this differentiation process. Work by Knapp et al. highlights the challenges in visualizing differentiation through time as well as identifying and labeling cell types [7]. Using their own data generated using CyTof (a related technique to scRNA-seq which allows for capture of 100-200 proteins per cell), they performed a functional cell type labeling strategy and connected this with molecular data, contrasting the two. Their unique analysis required an atypical strategy when assigning cell type, as their molecular data was often disconcordant with functional information. Their efforts are summarized in Fig. 2 - after collecting their molecular data and performing clustering via t-SNE, they overlaid functional differentiation data allowing for a comparison of molecular and functional states of their cells. Their main aim is to instill a sense that the molecular data does not capture the functional data, suggesting that our understanding of the system is limited. Some specific design choices were made, and are clear when compared to typical DR plots. First, no marks can be seen for individual cells - instead, a black outline is given for the density of the population on the plot (75% of all cells falling within the boundary, centered based on density). Overlaid on top of this boundary is a shaded density map for a given cell type as defined by functional experiments, with one boundary and density map pair for each cell type. Overall the lack of marks for cells in these plots does allow the viewer to focus on the distribution, driving home the idea that for most cell types, molecular data does not capture their differentiation potential. To further drive this point home, the authors perform hierarchical clustering in Fig. 3 on each cell type using distances between clusters from the t-SNE plots in Fig. 2. In this case, by abstracting the t-SNE plots and combining them with the functional data in a new way, the authors reiterate their conclusions in a less subjective way than subjectively judging the plots themselves.

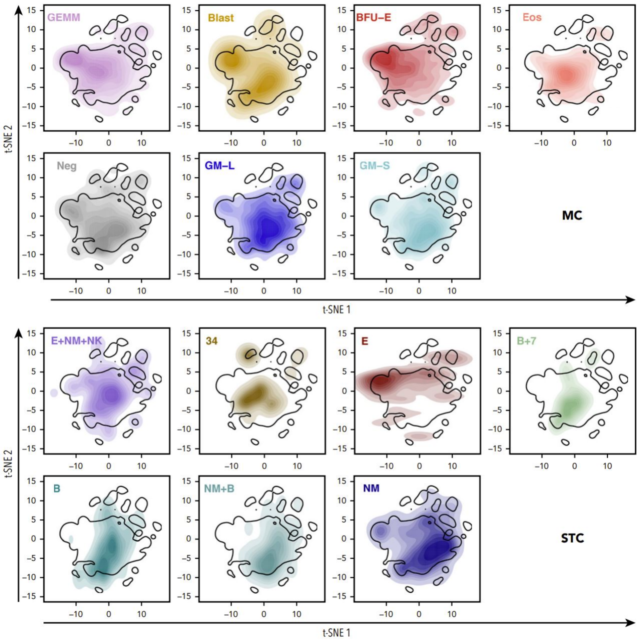


Fig. 2. t-SNE plots of bone marrow stem cells undergoing differentiation as analysed by CyTof. Black outline indicates the 75th percentile for cell density in the plot. For each duplicated t-SNE plot, a unique overlay is shown to indicate where cells of a given cell type have clustered.

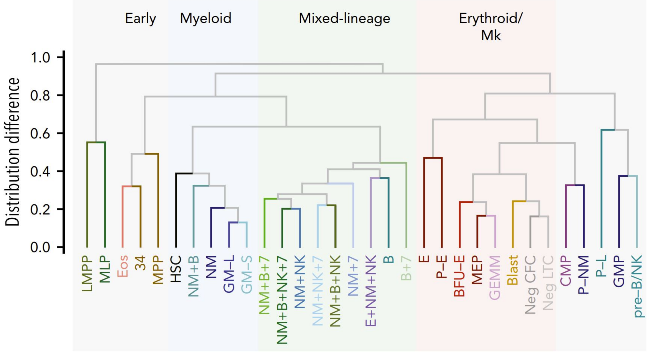


Fig. 3. Hierarchy of cell types and states as defined by Knapp et al. Cell types were functionally defined elsewhere in the original publication. Hierarchical clustering was performed using pairwise assessment of differences in the density distributions between all cell types.

Fig. 4 represents a novel way of showing pseudotime trajectory information that is often missing from scRNA-seq papers. Rather than collecting true-to-life time-course data, the authors of this paper leverage the existing heterogeneity of their system. Since blood differentiation is an active process, by analyzing cells at a single time point they collect several cells on a single biological trajectory. By performing an additional step of analysis, it is possible to infer a pseudo-time trajectory. Using the example of two such trajectories, the authors identify genes that dynamically change alongside the the trajectory (Fig. 4). To illustrate these changes, the authors use two approaches: heatmaps and scatter plots. Within the heatmaps, the authors place each gene of interest on its own row, with pseudo-time along the x axis. Accompanying these heatmaps are scatter plots indicating the same information, but for a single gene per plot. To the viewer, these two plots are not redundant, despite having the same genes and pseudo-time data on them. Each fills a distinct niche based on the depth required: the heat map gives a larger overview of the set of genes, giving information on 20 across the trajectory, while the scatterplot only provides 1 gene per plot. Still, while the scatterplot lacks the global view, it does give a much clearer indication of the distribution of cells and magnitude of change for each gene over time.

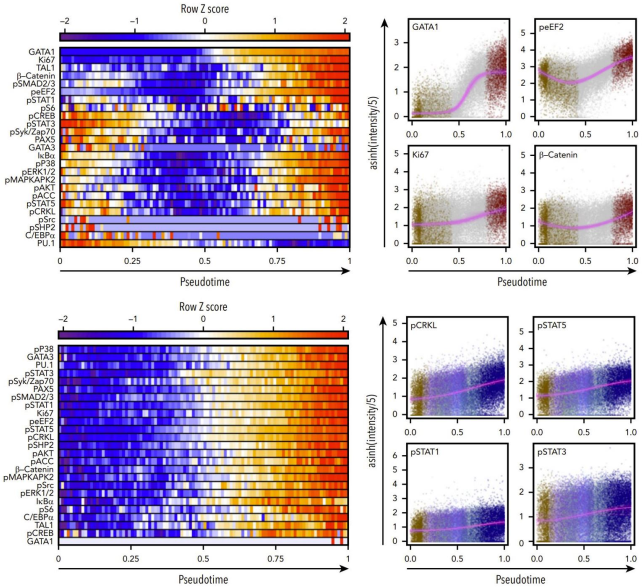


Fig. 4. Pseudo-time analysis of temporally dynamic genes. **Top and Bottom**: Two different pseudo-time analyses from the stem cell state to two distinct final cell states. **Left**: Heatmaps of genes correlating with specific pseudo-time trajectories from the stem cell state. **Right**: Scatter plots of select genes that are dynamic across the chosen pseudo-time trajectory in cells traveling that trajectory, with each point as a cell.

The last three years have seen a gold rush of competing papers aiming to achieve the most cells sequenced in a single publication. Often, these papers are enabled via advances in single cell sequencing technology or RNA isolation technology. However, as the number of cells and complexity of the dataset grows, it remains valuable to keep an eye on how the field decides to visualize the state of the art. One recent example from 2019 by Cao et al. highlights this trend [1]. Throughout the paper, the visualization standards of the field predominate. Still, some new approaches are taken, especially when zooming in on a specific subset of the paper.

When looking to specific figures in Cao et al., [1], the expected plots can be seen. In Fig. 5, we are given an overall summary of the massive dataset published by this group. Despite the size of the set, t-SNE clustering and the accompanying dot plot provide a good general overview of the dataset. By breaking down the clustered cells into day-specific t-SNE plots, they also provide a degree of temporal information - viewers are able to search and see approximately when their cell type of interest is first identified by looking at the dot plot and then cross referencing it to the smaller t-SNE sub-plots. Additionally, the dot plot provides a valuable resource for researchers interested in a specific cell type, as now there is a central location for identifying markers for your cell type of interest. One critique, however, is if this is really the best use of space given than for almost every single marker and cell type pair, there is a single strong signal. The usage of a dot plot as a visualization idiom would be ideal if expression was noisier or more of the markers were expressed widely or to varying degrees, but as most cells only express one of these markers highly the full plot does seem excessive.

Throughout their publication, Cao et al. perform pseudotime analysis. One way they visualize this is seen with Fig. 6. In this example, the authors pick out 6 key genes they identify as dynamic during a specific stage of development. Using colour (as seen throughout almost all of their plots), they encode which day these cells originated from. By plotting the relative expression versus pseudotime, as well as including the colour encoding, the viewer gets an idea of the dynamic nature of each of these genes. One key note is the large number of zero values, as is often the case in scRNA-seq experiments. For all 6 genes, a large number of cells do not seem to express these markers, while others have a much higher relative expression. This is valuable to see, as it gives the viewer both an indication of the noise in this data as well as the amount of missing values or dropout per gene. To remedy this, the authors add in a black line as an inference for expression patterns seen during this period in pseudotime. These plots can be directly contrasted to those in Fig. 4, where the goal is also to show pseudotime analysis versus target genes. Lacking from Cao et al. in this specific example is the same higher level heatmap of genes versus pseudotime.

One interesting visualization choice that was seen in Cao et al. was the inclusion of a 3D UMAP plot seen in Fig. 7. While the use of 2D t-SNE and UMAP is sprinkled throughout this publication, the authors decided to include a 3D plot for this one instance of trajectory inference. Without any prompting, a viewer might actually mistake this plot for a 2D UMAP if not for the faint lines behind the cells themselves. There is great difficulty in determining how the third dimension plays into the plot itself (for example, the bottom left cluster could be in any orientation given our single view). The contents of the plot itself are similar to others in the publication - the consistent use of colour as a channel to encode the day of origin for the cells is again seen here. Added too are cell type and trajectory labels to inform viewers of the nature of the clusters.

Finally, the authors take a novel approach to assigning expression values to cell clusters in Fig. 8. First, they generate a subsample of a set of cells of interest and re-cluster these cells. To this new plot, they then generate a set of overlays, where cells that express a given marker are labeled based on their level of expression. Critically, cells without any expression of the given marker are excluded from each of the plots. By performing this for each gene, the authors build a map of expression based on the clusters that is not heavily impacted by occlusion or is not too crowded by cells with zero expression. These plots then inform the final plot in the figure, where the overall cell states in the re-clustered group of cells is inferred using the individual gene markers.

One key take away that is well showcased here is the difficult balance between resolution and the scale of visualization. When looking at Fig. 5, we can see how the authors attempt to give an overview of their data via large-scale t-SNE of all cell types and time points. However, when more granularity is required, such as when performing trajectory analysis or looking at gene expression across pseudotime such as in Fig. 4 or Fig. 6.

4 Methods and Tools

This dataset has not been previously analyzed by BK nor KM, nor by any other groups. It was released in late 2022 as part of a Kaggle competition [2]. We chose to use the two most common packages for scRNA-seq analysis are Scanpy and Seurat [5, 21]. These tools are considered the gold-standard for initial exploratory analysis of single cell data, and have been continually expanded over the last half decade [4, 14, 17]. These packages are primarily are geared towards performing statistical analysis and handling the data, but also have dedicated components for plotting and visualizing the resulting data after normalization and cleaning. We chose to use both as functionally they provide the same core set of idioms and flexibility with design choices. For more specialized analysis, we chose to implement specific tools such as monocle3 for trajectory inference and PAGA for graph abstraction [1, 22]. While Seurat and Scanpy do provide some basic tools for trajectory inference, they fall short when compared to dedicated packages such as Monocle3. Additionally, Seurat and Scanpy completely lack the ability to perform graph abstraction on their resulting DR plots, making a dedicated package such as PAGA a necessity if such visualization is required. Monocle3 was chosen specific due to its performance in trajectory inference as well as its relatively low computational requirement when recently compared to other inference algorithms [23].

Common tools such as Scanpy and Seurat use a common set of idioms for visualization. Within this set of tools, we set out to explore which best achieved our tasks, and to explore the trade offs between different idioms. Typically, scRNA-seq analysis heavily relies on the same set of idioms: linear and non-linear dimensionality reduction, heatmaps, scatterplots, histograms, and connected graphs. This primarily is due to the nature of the data and the types of conclusions that are drawn from the data. Given the data is typically quantitative count data, visual idioms that allow for comparison of counts dominate. Given the size of the datasets, dimensionality reduction naturally fits given its ability to allow visualize large complex datasets. Ultimately, as our goal was to evaluate available tools, we narrowed our scope to investigating visualization idioms within these packages.

5 Analysis

5.1 Quality control

5.2 Global view

5.2.1 Dimensionality Reduction

Initially, we sought to characterize the gene expression dataset in its entirety with the most commonly utilized visual encodings. DR methods form the bulk of these methods through their ability to observe global trends with respect to changes in a clustered attribute, the assigned cell type labels. In addition, the flat table was stratified by the day attribute. Constructed using Scanpy, these DR methods were called with default parameters. Notably, for the non-linear DR methods (UMAP and t-SNE), Scanpy performs an initial PCA reduction in order to maintain scalability. Many single-cell analyses perform PCA as a pre-processing step, and both Seurat and Scanpy provide visual encodings to ensure that the reduction method is appropriate.

Fig. 10 depicts common encodings that are used to validate the use of PCA as a pre-processing step. The left plot shows a heatmap of the principal components when grouped by the cell type attribute along the y-axis (Fig. 10). We can see that the first several components appear to be sufficient to capture a majority of the attribute-specific variation that exists in the expression table. The right plot features a scatterplot that shows the amount of variation that each principal component captures (Fig. 10). Interestingly, there is a large drop-off in the standard deviation explained after the first principal component, which may suggest that this component captures the relative majority of gene expression variation. Moreover, after approximately 15 principal components, additional components capture relatively negligible variation in the dataset (Fig. 10). Overall, the two visual encodings in Fig. 10 are able to confirm that such a pre-processing step is justified for downstream analyses, and based on these results, at least 20 principal components were used in all remaining analyses when working with the reduced gene expression dataset.

Fig. 11 depicts the reduced gene expression data as a scatterplot at each time point following DR with PCA, t-SNE, or UMAP. DR with PCA was unable to capture any separation with respect to the clustered attribute, and there were no obvious patterns over the global cell population Fig. 11. In contrast, both non-linear DR methods were able to capture global trends, which moderately aligned with the assigned cell type labels. There is not clear separation with respect to the clustered cell type labels, and this makes it difficult to view the boundaries for the different groupings. To address this, the scatterplot of the UMAP embeddings were further faceted by day and cell type in Fig. 12. In particular, the two-dimensional embeddings are plotted for the entire cell population, and each cell type is highlighted in turn using its respective color channel (Fig. 12). This superimposition provides a much clearer view on the distribution of different cell types, and how those change depending on the day. While some cell types such as erythrocyte progenitors, neutrophil progenitors, and hematopoietic

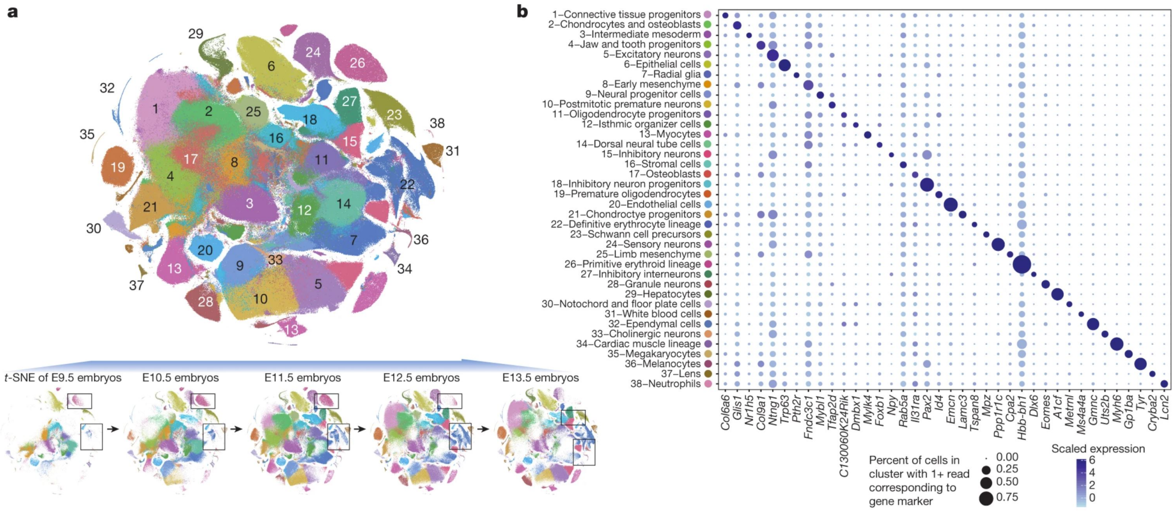


Fig. 5. Overview of the 2072011 cells sequenced from early mouse embryonic development by [1]. **Left**: t-SNE clustering of single cell transcriptomes generated from this study. Clusters are annotated and coloured according to the adjacent dot plot. Individual clustering was performed and shown below for each of the embryonic days (EX.5). **Right**: Dot plot indicating specific genes that are selectively expressed in a given cell type. The size of the dot indicates the percentage of cells in that type that express the marker, and the colour encodes the average expression of that gene in that cell type.

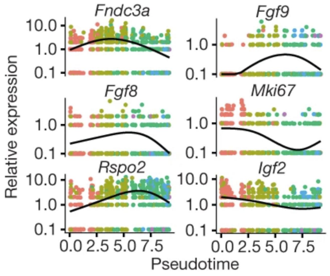


Fig. 6. Scatter plot of specific genes that are dynamically expressed during apical ectodermal ridge development. Expression is plotted versus pseudotime for each cell. Colour encodes which time point the cell originated from.

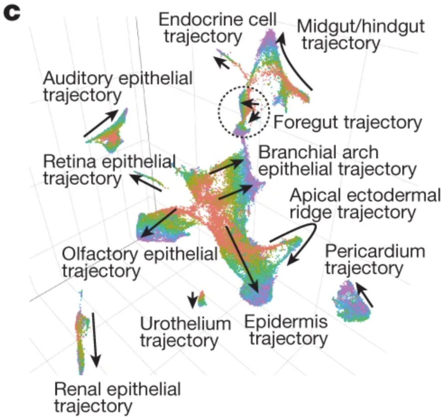


Fig. 7. 3-dimensional UMAP of epithelial sub-trajectories. Colour encodes which time point the cell originated from. Labels are applied according to trajectory inference and specific marker gene expression.

stem cells have relatively tight groupings over the three time points, the locations of other cell type labels appear quite dispersed over the cell population (Fig. 12).

One of the main advantages of the dimensionality reduction visual encodings were that they can capture global trends of the dataset, which is particularly useful for evaluating attribute-specific clustering. In terms of domain-specific findings, the superimposition in Fig. 12 was particularly helpful in understanding the overall distribution of different cell types. Generally, cell types were grouped together, but there were a few cell types such as monocyte progenitors, B-cell progenitors, and megakaryocyte progenitors that were quite scattered throughout the overall population. In the case of B-cell and monocyte progenitors, the relatively small number of collected cells confounds whether their scattered distribution is meaningful, whereas with megakaryocyte progenitors, the grouping did get tighter over time. As such, through these global visual encodings, we concluded that while there was not clear separation with respect to the assigned cell type labels, they may be appropriate as a very high-level, conservative grouping attribute. While applying algorithms such as Leiden clustering [19] were able to find approximately 20 clusters, it is highly unlikely that such clusters are biologically meaningful.

The main weaknesses of these scatterplots following DR methods is that the interpretations are strictly qualitative. The main insight that these encodings are able to provide is whether there is clear separation of the clustered attribute. Moreover, readers are not able to precisely compare such the relative separation of clusters across different time points. Another limitation is that plotting the two-dimensional embeddings of non-linear dimensionality reduction methods require a high cognitive load when deriving interpretations. This is because DR methods such as t-SNE and UMAP do not preserve meaningful distances between different clusters. So while their non-linear reduction methods may be able to effectively capture local structures in the data, caution must be applied to avoid assigning any meaning to the extent of any cluster-specific separation.

5.2.2 Trajectory Inference

Further extending DR methods are a class of more recent statistical methods known as trajectory inference. Many trajectory inference methods operate by computing a pseudo-temporal orderings over the cell population, which can be used to assign an ordering to the different cell type clusters. Having such an ordering relationship between cell type clusters can be used to explain possible patterns that may arise from the gene expression levels of different cells. As such, the Monocle3

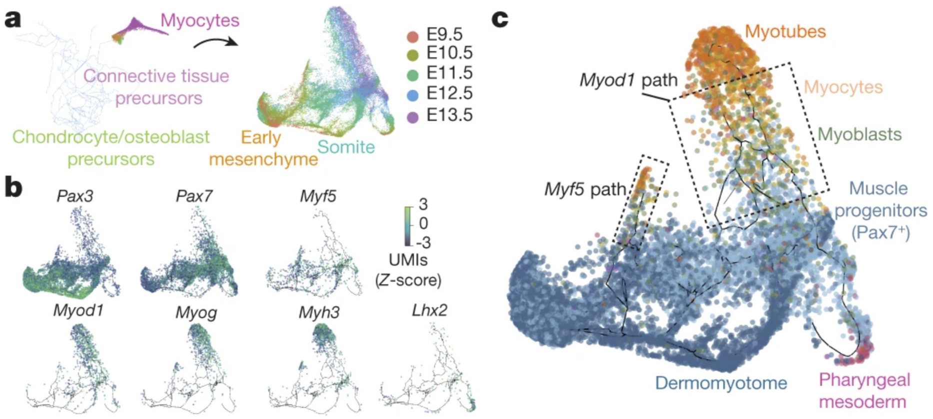


Fig. 8. Sub-trajectory analysis of muscle cell development in the early embryo. **A**: Cells with myocyte potential (muscle cell potential) were computationally isolated and reclustered, labeling each cell with a colour corresponding to its day of origin and adding an inferred trajectory to the clusters. **B**: Labeling expression of specific genes involved in controlling myocyte development in the myocyte trajectory from **A**. Cells with no expression of a given gene are not included in the plot for that gene. **C**: Cell type inference and annotation from the markers shown in **B**.

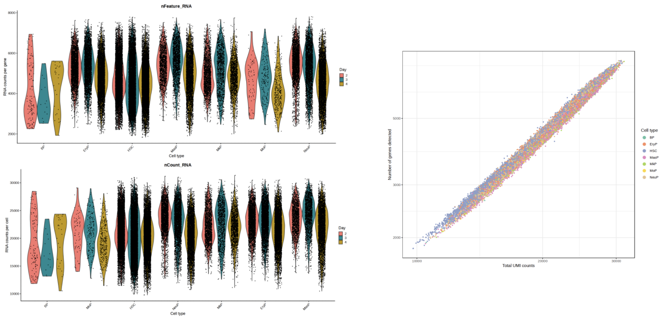


Fig. 9. t-SNE dimensionality reduction of cells across days 2, 3, and 4. All cells are indicted on each plot by either a grey dot or a coloured dot. For each cell type, a unique plot is made for each day, with cells of a given cell type coloured for reference.

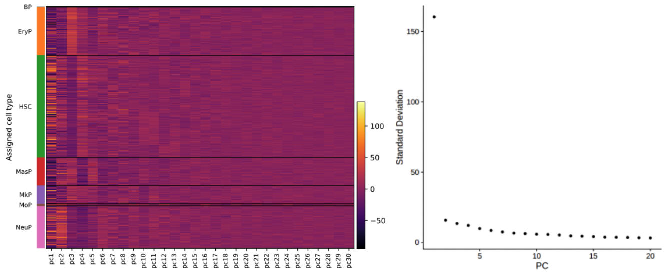


Fig. 10. UMAP embeddings faceted by day and cell type. Each row depicts the superimposition of a cell type in the context of the rest of the population which is shown in gray.

software [1] was used to construct single-cell trajectories through the inferred pseudotime.

Fig. 13 depicts the inferred trajectories with respect to the UMAP embeddings of gene expression. Interestingly, as identified by the numbered light gray circles, performing trajectory inference analysis finds 15, 11, and 17 different ”cell fates”, or outcomes, of the gene expression trajectory for days 2, 3, and 4, respectively Fig. 13. This finding may provide evidence suggesting that further breakdown of the cell type labelling may be justified.

Hematopoietic stem cells were specified as the starting attribute-specific cluster on which to start the pseudotime inference from because they are still capable of differentiating into any mature blood cell type. Fig. 14 shows a scatterplot of the UMAP embeddings at day 2, when augmented with the inferred trajectories. The color channel is used to represent the continuous pseudo-temporal ordering that was inferred. The HSC cell type is shown at the bottom of the scatterplot, and the gene expression pattern suggests a differentiation trajectory where it branches out to the cell type-specific progenitors over time (Fig. 14).

Overall, the trajectory inference visual encodings are capable of provide additional directionality to the compared to the DR visualizations alone. A major advantage that comes with this pseudo-temporal ordering is that it can infer branching points at which cell differentiation trajectories may split into different mature blood cell types. It provides a way to interpret changes in attribute-specific clustering in a continuous manner, which more accurately models the dynamic change in gene expression levels within and between the different cell type clusters. This is a particularly suitable dataset on which to apply trajectory inference analysis due to the focus on the very early stages of differentiation where many cells are in the midst of responding to factors that influence their eventual differentiation into a mature blood cell type. Moreover, the branching structure of the inferred trajectories may help explain the relative scatter of certain cell types such as megakaryocyte progenitor cells across the population.

That being said, trajectory inference methods suffer from similar disadvantages that DR scatterplot visualizations face, in that it is difficult to compare the inferred trajectories across different time points.

5.2.3 Networks

Another visual encoding that may be able to capture global trends in attribute-specific clusters is Partition-based graph abstraction (PAGA). By generating a topology-preserving map of individual cells according to their cell type label, PAGA graphs can view the aggregated global topology in an interpretable manner.

Fig. 15 depicts the the PAGA graph for the cell type labels at each time point, which represent the course-grained connnectivity structures of complex manifolds. This abstracted graph encodes confidence scores on the presence of connections by edge weight, which can be interpreted similarly to typical bootstrapping methods. The aggregated connectivity edges between cell type clusters represents an ensemble of single-cell paths according to their respective gene expression levels, which allows use to qualitatively view the degree of connectivity between cell types across days. Interestingly, the neutrophil progenitor cells initially show relatively thick edge weights with neighbouring progenitor cells initially, but in the two later time points, the connectivity decreases (Fig. 15). This decrease may suggest that changes in gene expression may be stabilizing for a distinct subset of genes that characterize mature neutrocyte cell functionality. The observation that the PAGA graphs are fully connected at every time point likely reflects the fact that cell

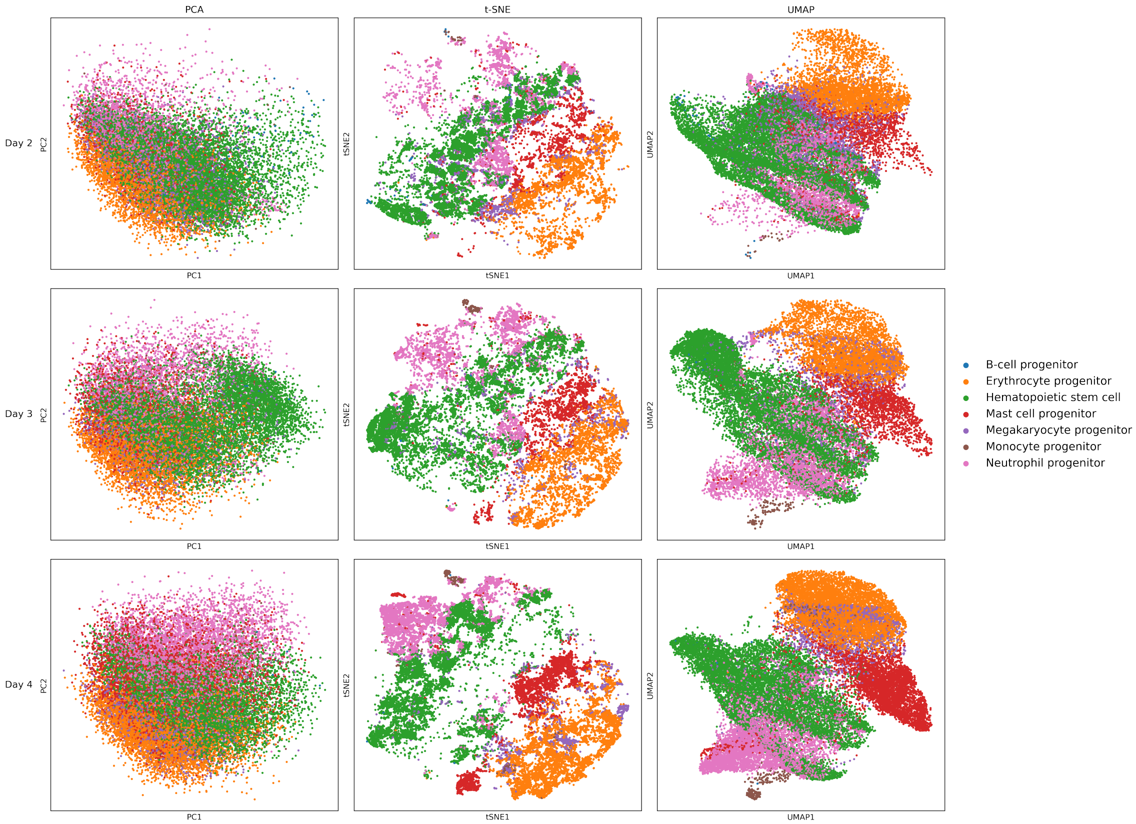


Fig. 11. Comparing scatterplots of gene expression data at each time point following dimensionality reduction with PCA, t-SNE, or UMAP. The color channel depicts the assigned cell type. **Left column**: linear dimensionality reduction with PCA. **Middle column**: non-linear dimensionality reduction with t-SNE. **Right column**: scatterplot of two-dimensional UMAP embeddings.

type clusters are fully differentiated into their mature blood cell types (Fig. 15).

While PAGA graphs can depict global topological structure of attribute-specific clusters without placing a high cognitive load on the reader. The aggregation with respect to cell type labels was necessary in order to capture the global trends in this dataset. Additional network-based visual encodings were evaluated, but they focused on treating single cells or genes as nodes, which incurred significant clutter and occlusion to the point that deriving any patterns became practically impossible.

A disadvantage of PAGA graphs is that the aggregation may lose a significant amount of variation that exists within cell type clusters. Moreover, the implementation of PAGA graphs provided by Scanpy does not permit the depiction of cell type labels in a user-friendly manner, as there is no in-built argument that moves the cluster legend to the side of the plot.

5.2.4 Optimal transport

While many trajectory methods infer a pseudotime that can subsequently be used to order the cells along a trajectory based on the continuous change in gene expression levels, these methods are still limited in that they cannot handle such inferences across multiple time points. Thus, while the psuedotime-based methods can capture insights with respect to the trajectory of the dynamic changes of gene expression between different cell types, they are not able to capture the true essence of trajectory inference, which is to infer the differentiation trajectories of individual cells across time.

A recent method has been developed that performs the trajectory inference over a time-series dataset of gene expression at single-cell resolution called WaddingtonOT [15]. Using the assigned cell type labels as clusters for each time point, this approach relies on optimal transport and constructs trajectories of cell differentiation by effectively stitching together individual cells across time based on similarities in gene expression profile. As a result, we expected this method to be able to address many of the limitations of DR visual encodings and pseudotime-based trajectory inference methods by appropriately considering the time-series aspect of this dataset.

Unfortunately, we were unable to successfully apply the software to the current dataset, as the inference process failed upon forming linking maps that function to connect individual cells across different time points.

5.3 Subset view

5.3.1 Top ranked genes as a subset

Many visual encodings in single-cell analysis circumvent having to deal with the high dimensionality of the dataset by extracting a subset of genes that were found to be differentially expressed. Differential expression is typically defined in terms of a statistically significant difference in expression levels between two groups – for example, one cell type versus the rest of the population. Filtering to only include specific attributes (genes) is based on the rationale that expression datasets are typically sparse, and a given cell type can oftentimes be defined by the expression – or lack thereof – of a subset of genes in a cell's DNA. Regardless of whether the identified subset of genes is associated with an underlying biological mechanism, this approach can effectively identify signals in gene expression that have led to many interesting findings at the level of the biological system. Visualization has been shown to be an essential component in this process of selecting for genes that are exhibit relatively high or low expression for a particular cluster of cells compared to the rest of the population.

The most straightforward practice for identifying differential genes for a given cell type involves ranking the aggregated expression levels

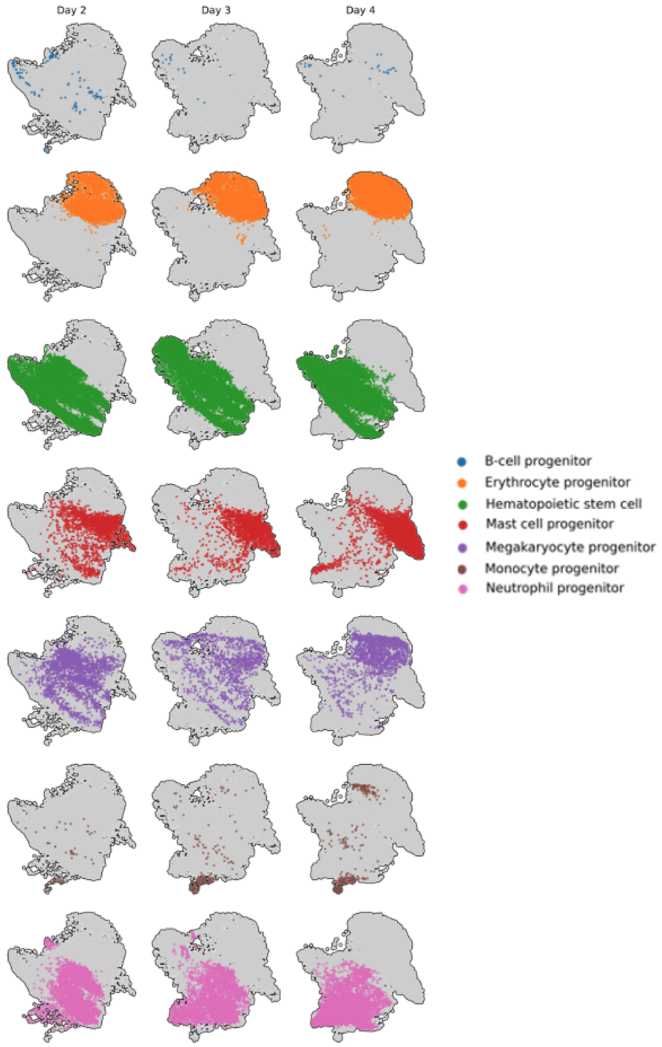


Fig. 12. Visualization strategies to analyze the appropriateness of PCA as a pre-processing step for downstream analysis. **Left**: A heatmap of the principle components where the color channel is used to depict the PC values. Cell types are grouped along the y-axis. **Right**: An elbow scatter plot depicting the amount of variance that each principal component is able to explain.

of each gene over the subpopulation of cells. Once the ranking has been computed, we can simply select the *n* highest-ranked genes to use as markers to define that cell type.

One of the most popular visualization idioms for this subset view is the heatmap. The heatmap in Fig. 16 depicts the top 3 ranked genes in terms of expression level for each cell type at the different time points. The cells are grouped by their cell type labels along the y-axis, and the names of the genes are provided on the bottom, with an annotation that identifies cases where the same gene was identified as top 3 expression level at multiple time points (Fig. 16). Clear cell type-specific patterns emerge with this visualization idiom, as the genes with highest ranked expression appear darker (Fig. 16). In particular, EryP, HSC, and NeuP cell types had ranked genes that were distinct from all of the other cell types (Fig. 16). In addition, a majority of the idenitified genes were identified at multiple time points as denoted by the color channel highlighting the gene names, which suggests that such genetic modules may be characteristic of the defined cell type clusters.

While heatmaps can be quite information dense, one of their advantages is that they can capture the variability within and between cell type labels for the subset of genes visualized. The act of grouping such genes by cell type also has the benefit of conveying the relative sizes of the cell type clusters, and this insight may act to implicitly instill confidence to the reader with respect to whether the ranked genes should be trusted. For example, given the relatively large population of EryP cells, the ranked genes identified likely display a distinct enough pattern to justify the use of those genes in downstream analyses (Fig. 16). In contrast, the small sample sizes for the BP and MoP cell types are likely to convey a level of uncertainty due to their susceptibility of sampling bias, which may would at least serve to caution users from assigning too much weight, or trust, on the identified genes.

One of the main disadvantages of the heatmap visualization idiom in the context of this dataset is that it is difficult to make quantitative comparisons across time points. Moreover, while using the color channel to depict expression level is effective for qualitative identification of gene expression patterns, the human eye (and brain) is not sensitive enough to detect small changes in expression level.

Scanpy offers a similar visualisation idiom where instead of depicting individual cells as a heatmap, it aggregates the expression values and depicts the aggregated value as a dot. The size of the dot corresponds to the fraction of cells in the cell type cluster that detected the respective genes, and the color channel is used to represent the aggregated expression value. This encoding is shown in Fig. 17 for the same top 3 ranked genes given each cell type.

The primary advantage of this visual encoding is that there is far less cognitive burden compared to the heatmap idiom. Depicting the aggregated expression values as dots may be more conducive to identifying patterns in the dataset, particularly for the rare cell types, such as BP and MoP (Fig. 17). The relative size of the dot plot suggests that the top ranked genes for the BP cell type may be distinct compared to the other cell types for days 3 and 4 (Fig. 17). Similarly, the dot sizes may also suggest that the top ranked genes are meaningful for the MoP cell type at day 2 (Fig. 17). Thus, the additional information encoded into the size of the dots visual idiom may lead to a different interpretation with respect to the reliability of the markers identified for the rare cell type groups. The heatmap was unable to convey such patterns due to the small sample size.

The dot plot visualization idiom has a couple disadvantages in that the color channel and size of the dot is used to convey important information, despite the fact that they are two channels that the human eye cannot quantitatively assess with precision. As such, this idiom is likely better served to focus on identifying high-level patterns at the cell type cluster level.

We assessed one other visualization idiom for the top 3 ranked genes with respect to each cell type. Fig. 18 is similar to the previous two idioms but it instead encodes a violin plot to represent the expression information for each cell type group. In a sense, the violin plot aggregates the single cell expression information to attenuate the amount of cognitive load on the reader, but unlike Fig. 17, the violin distribution that it encodes retains much of the variation information within each cell type.

This particular idiom can depict within-cluster skews of expression information in a very interpretable manner compared to the heatmap idiom, while also being able to qualitatively identify high-level patterns at the cell type level. One interesting insight is that many of the top ranked genes are not expressed in every cell in the cell type cluster, which would pose challenges if trying to utilize the genes as a marker for its respective cell type. This particular insight may guide the single-cell analysis toward more robust methods when identifying markers that are truly representative of the cell type of interest.

5.3.2 Differential expression methods to identify a subset

While identification of the top-ranked genes for each cell type is a quick strategy that can effectively extract signals from the high-dimensional dataset, there are generally two major disadvantages. First, there is not guarantee that selecting the top-ranked genes for one cell type will be mutually exclusive for another cell type. For example, ”housekeeping” genes refer to genes that are typically expressed by all cell types because they play a vital role for the maintenance of basic cellular function. Second, consideration of only highly expressed genes only views the task from a single angle when in fact the low expression of genes can

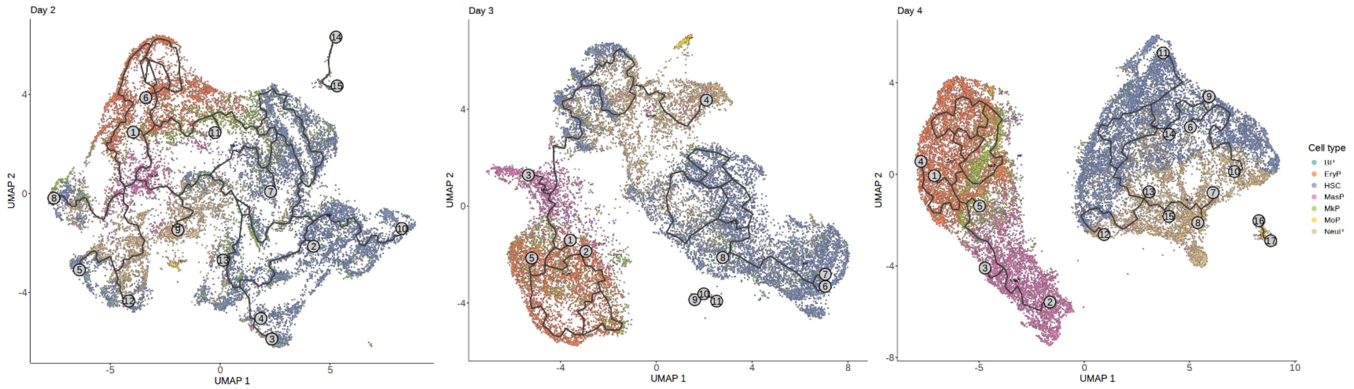


Fig. 13. UMAP dimensionality reduction plots with overlaid inferred pseudotime trajectories. Individual cells are labeled with cell type information provided from the original dataset. Inferred trajectories were generated using Monocle3. Cells were separated based on their day of origin and plotted for each day. Individual clusters or cell states are annotated with number labels according to the output of Monocle3.

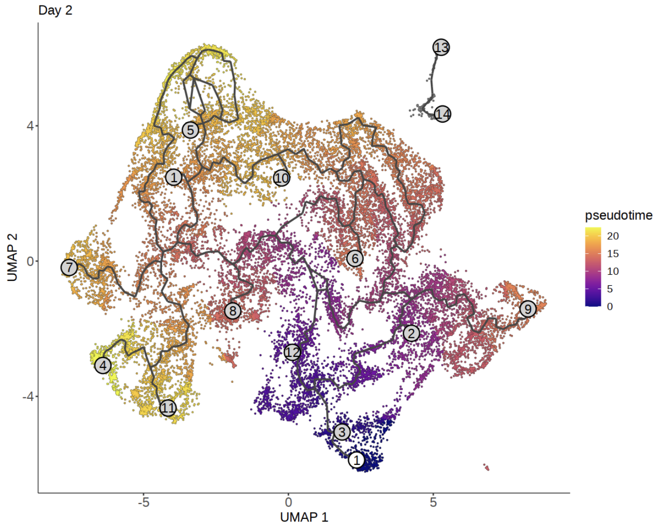


Fig. 14. test1 Individual pseudotime embedding of cells from day 2, as plotted in Fig. 13. Cells are coloured according to their position in pseudotime, with inferred trajectories overlaid via black lines. Individual clusters or cell states are annotated with number labels according to the output of Monocle3.

serve as an equally effective marker for a particular cell type.

Fig. 19 may point out some of the limitations of simply selecting the top-ranked genes. Despite NeuP cells demonstrating one of the most distinct expression patterns among the assigned cell types, when comparing the distribution of expression levels for NeuP cells against the rest of the cell population, we see there is some degree of overlap between the side-by-side distributions (Fig. 19). As a result, if any of the top-ranked markers were used as marker genes in practice, they would be unable to cleanly discern NeuP cell types.

The main hallmark of an effective marker gene is that it can be reliably used to distinguish a cell type of interest from every other cell type in the population. An example of a set of well-known markers genes is for pluripotent stem cells. In 2006, it was demonstrated that introduction of the Myc, Oct3/4, Sox2 and Klf4 genes were sufficient convert a mature cell back to a pluripotent stem cell [18]. Given that the collective expression of these genes is well-documented in the context of pluripotent stem cells, it represents a reliable marker that can distinguish pluripotent stem cells from other cell types.

A popular method of identifying differentially expressed genes that relies heavily on visualization is referred to as a violin plot. There are two key principles underlying the application of conducting an unbiased global approach at identifying marker genes. The first principle involves focusing on changes in gene expression that are statistically significant. The second principle is a rule of thumb that the largest changes in gene expression are typically the most likely to be biologically relevant.

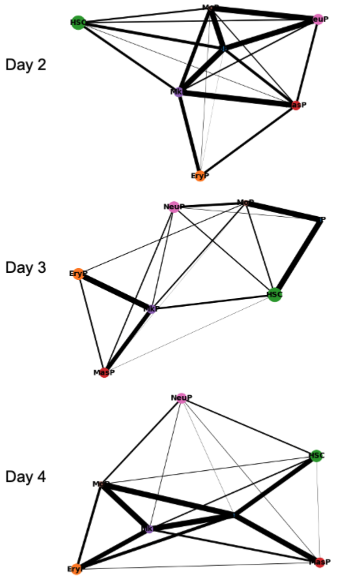


Fig. 15. Graph-based abstraction cells shown in Fig. 14. Graphical representations were generated using PAGA, with each node representing a cell type as given by the original dataset. Edge weight is encoded by the density of each edge, which represents the degree of similarity between nodes. Graphs were generated for each day of origin for cells from the original dataset.

Focusing on a single cell type grouping, for each gene we can split the entire cell population into 2 conditions: the cell type group of cells and the rest of the population. We can then apply a statistical test such as t-test to calculate the p-value with respect to each gene. Note that genes that were completely undetected by the cell type population of interest were excluded from the statistical test. At this point, we will have a p-value for every gene that compares the cell type group population against the rest of the cell population, and to account for the multiple comparisons problem we can subject the p-values to a Bonferroni correction. The transformed p-values are the first element of the volcano plot, where higher values suggest that the difference between the two

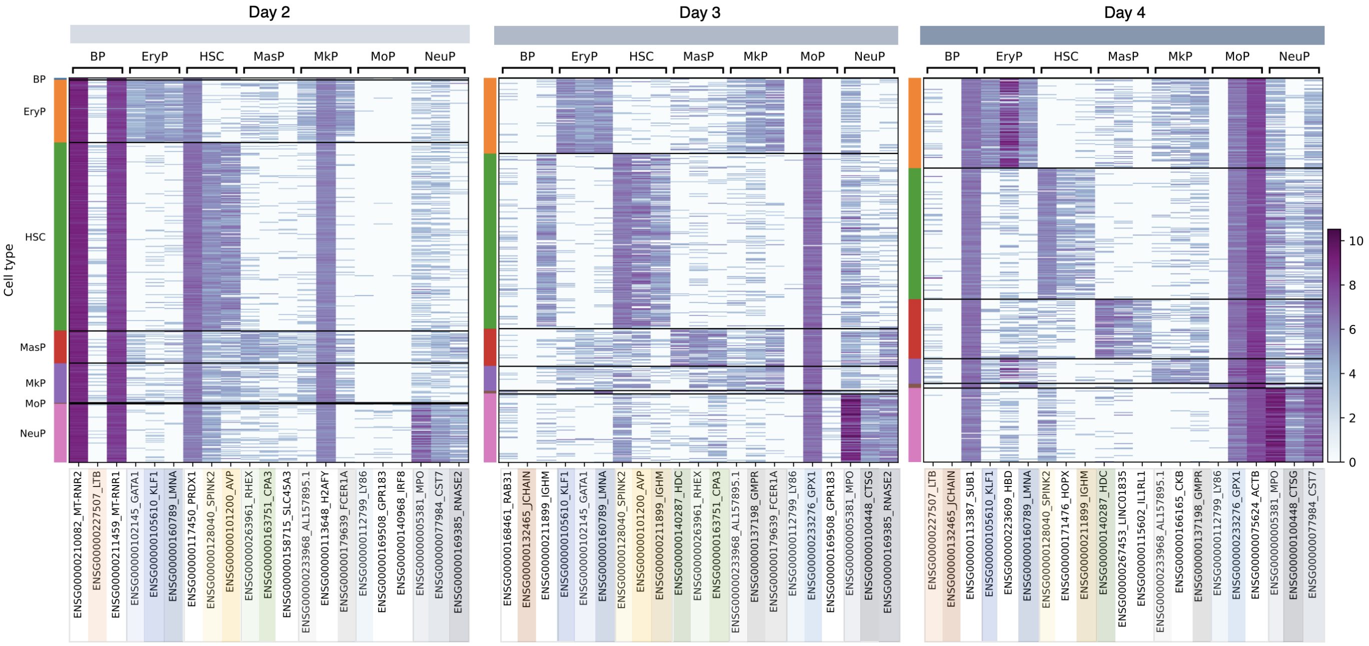


Fig. 16. Heatmap of the top 3 genes demarcating each cell type for days 2, 3 and 4. Cells are grouped according to their cell type label along the y-axis, with gene IDs listed at the bottom.

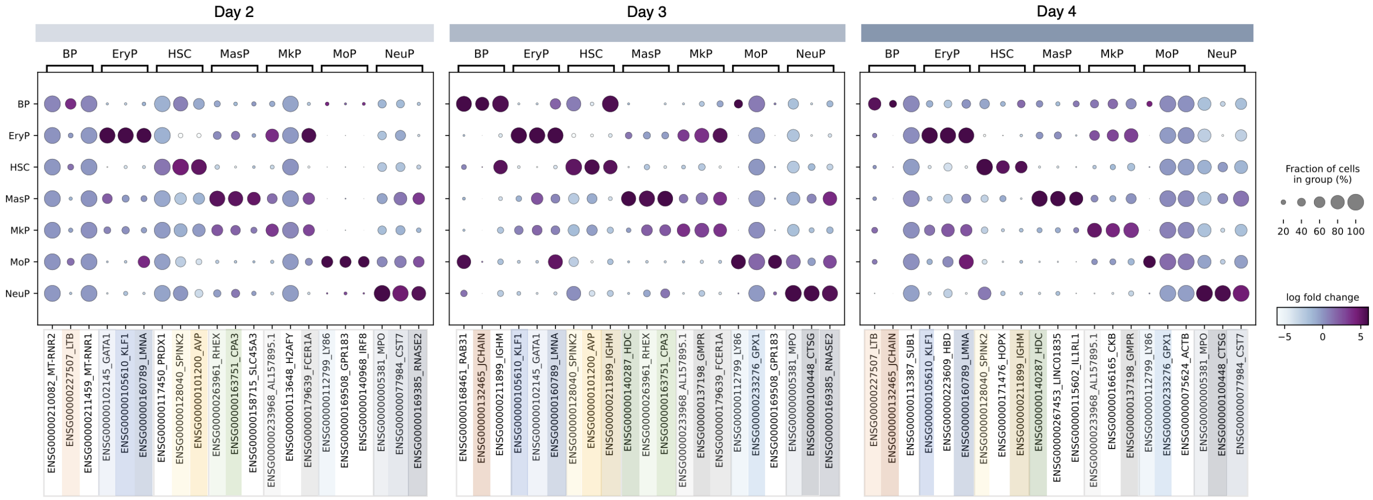


Fig. 17. Dot plot of the top 3 genes demarcating each cell type for days 2, 3 and 4. Cells are grouped according to their cell type label along the y-axis, with gene IDs listed at the bottom. For each dot mark, both size and colour are used as channels to encode information. The size of each dot encodes the percentage of cells in that cell type that express a given marker, and the colour of each dot represents the log-transformed fold-change versus the mean for each gene.

groups is more significant. The second element involves calculating the fold changes of gene expression between the two groups of cells.

We can plot the two elements for each gene and day to obtain the final volcano plots shown in (Fig. 20). In Fig. 20, each row corresponds to the time point at which sequencing took place, and each column corresponds to a cell type label. Given the visual encoding, we can define a threshold for the fold change and transformed p-value to determine genes that are differentially expressed. Fig. 20 shows the genes that are under-expressed in orange and over-expressed in blue, with the non-significant genes depicted as gray dots. For each panel in Fig. 20, the two elements were combine to create a score, and the top 5 differentially expressed genes were selected as marker genes for that cell type at that time point. Labels are provided for the top differentially expressed genes, and the size of the point is also slightly increased (Fig. 20).

This volcano plot is able to visualize the process of filtering out genes that are not differentially expressed. It can also capture the magnitude of the differential expression, which users can consider when determining the most appropriate threshold for to answer their specific question.

The advantage is of this visualization idiom is that it captures information with respect to every attribute in the high-dimensional dataset. The superimposition of deferentially expressed genes encoded by the color channel reduces the cognitive load by a large margin.

In the context of this dataset, however, there is no a clear separation between under-expressed and over-expressed genes. In fact, a majority of the genes are have a fold change close to zero, which may be a result of the similarities in gene expression profiles for cells in the early

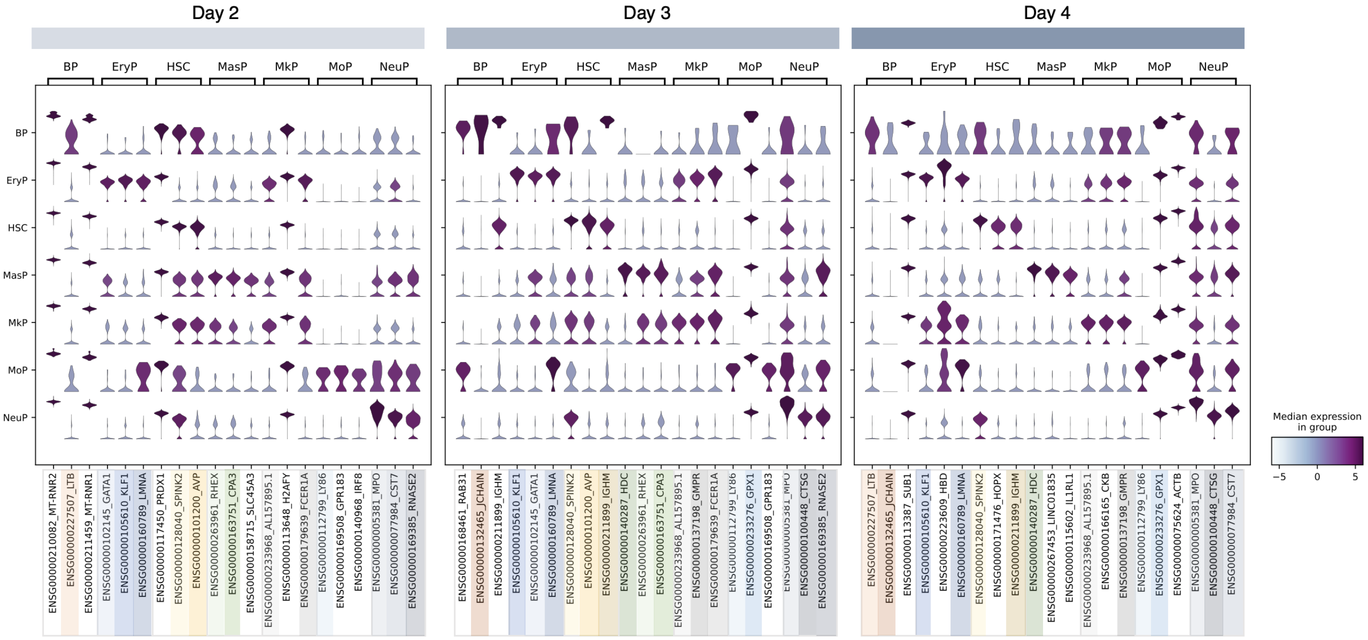


Fig. 18. Violin plot of the top 3 genes demarcating each cell type for days 2, 3 and 4. Cells are grouped according to their cell type label along the y-axis, with gene IDs listed at the bottom. For each violin mark, both shape and colour are used as channels to encode information. The shape of each violin encodes the distribution of expression for that cell type and gene. The colour of each violin represents the median expression of the indicated distribution.

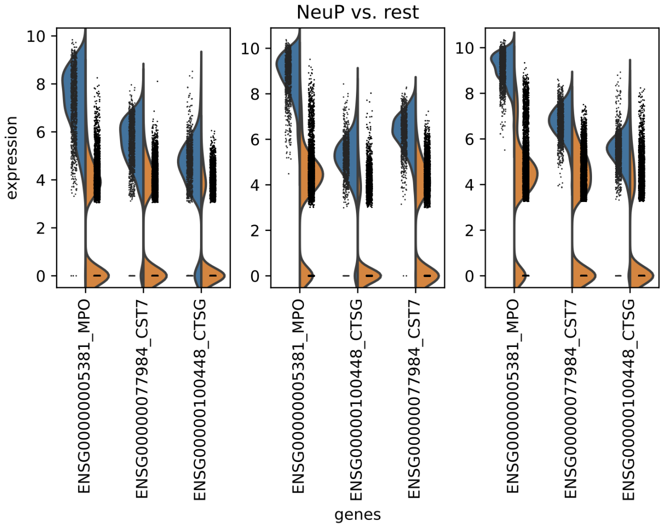


Fig. 19. Split violin plots of genes highly enriched in the NeuP cell type. Expression of these genes is compared with NeuP cells versus all other cells in the dataset. Distributions are represented via violin outlines, with individual cells overlaid for each group and gene.

differentiation stages. A peculiar observation with respect to this plot is that for the EryP, HSC, MasP, MkP, and NeuP cell types, the bulk of non-significant cells depicted in gray have a fold-change of nearly zero but are statistically significant (Fig. 20). The disconnect between fold change and transformed p-value may point towards issues that assumptions in the statistical test may be violated, or that the mean aggregation of gene expression values is not appropriate. Non-Gaussian variation in gene expression levels within each cell type cluster could explain either of these possible issues, and it could be solved by further breaking down the assignment of cell type labels into more specific subsets.

Fig. 21 shows an additional visualization idiom that can be used to assess the ability of the identified marker genes to discern the cell type of interest. It is able to capture more information than Fig. 19 by visualizing the distribution with respect to each cell type. Unfortunately, the differentially expressed genes do not appear to be an effective marker strategy, as there is overlap between the distributions for the cell type group and the rest of the population for each of the marker genes selected. While there are trends that could be used to distinguish the EryP cell population at this distributional level (Fig. 21), the degree of overlap is substantial which would prevent it from picking out most of the cell type of interest.

5.4 Time-series view

An element that makes the selected dataset interesting is that molecular information was collected at multiple time points. This application of visual encodings for single-cell analysis of time-series datasets is not well-established. That said, there was a need to characterize the existing softwares in terms of the support that they provide for visualizing such time-series data. The main limitation associated with the aforementioned visual encodings is that they could only make comparisons across time points in a qualitative manner. While such qualitative observations could still provide interesting insights with respect to the cell type labels, quantifying the dynamic change in gene expression levels could facilitate a deeper understanding of the dynamic change of gene expression as cells are in the early stages of differentiation.

It was found that there was a lack of support for in-built to provide this quantitative information. However, the most frequently applied encoding involves viewing a single gene at a time. By grouping the cells according to the cell type labels, a gene of interest can be aggregated and subsequently plotted across time as a line plot. To illustrate this approach, 20 of the most-differentially expressed genes were identified across the global cell population. Fig. 21 Shows the distribution of these variable global genes when visualized as a scatterplot according to the standardized variance against the average expression.

Using the differentially expressed genes over the global population, time-series line plots were visualized (Fig. 23). For each gene, the expression values are aggregated by cell type. As such, we can see trends of differentially expressed genes over the 3-day experiment.

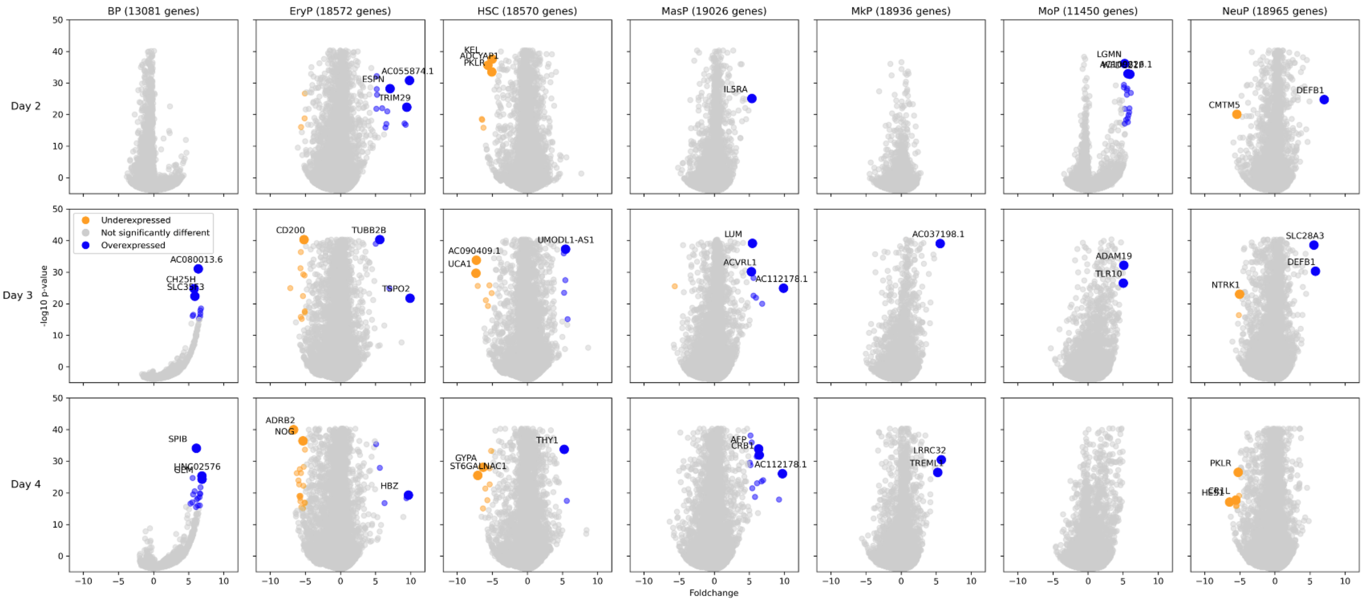


Fig. 20. Volcano plots comparing each individual cell type to all other genes across all three time points. Genes that are statistically significant are coloured (blue for overexpression, orange for underexpression). Statistical significance was generated by performing a two-tailed t-test followed by false discovery rate correction via Bonferroni method (α of 0.05 and 0.15 respectively).

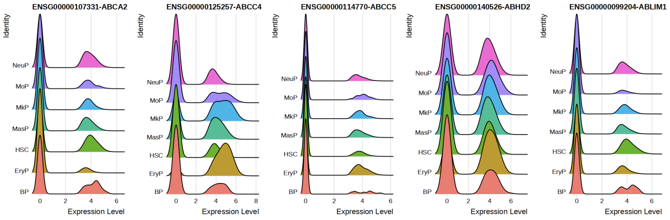


Fig. 21. Histogram of individual genes enriched in EryP cells. Each gene has the distribution of expression for each cell type plotted in a vertical stack. Cell type identity is encoded via the colour of each of the given histograms.

While there was little variation amongst the cell type labels for some genes, others were elevated or much lower for certain cell type groups (Fig. 23). In addition, the time-series analysis was able to capture the trending direction of gene expression change.

One thing to note is how there appears to be relatively more fluctuation in gene expression levels for the BP cell population. This is likely a reflection of the small sample size, combined with the scattered distribution among the overall population that was observed with the DR visual encodings. Overall, at the level of individual genes, this visual encoding can provide insight to the change in gene expression levels over time.

The main limitation with this approach is that only single genes can be assessed. That being said, the power that this visual encoding is able to provide has a lot of support from an interpretation standpoint, as scatterplots and line plots are able to convey information through channels that are easy to interpret in a quantitative manner. This may suggest that this limitation is more concerned with methods to identify the suitable marker genes. Yet it does raise a possible niche for the development of new visual encodings that are able to quantitatively capture dynamic changes in global gene expression levels over time.

5.5 Protein information

While transcriptomic information is the primary modality at which single cells are generally characterized, more recent technological advancements have allowed scientists to simultaneously unravel additional levels of information within cells, including protein levels. This has led to a new era of single-cell analysis, and CITE-seq is one of the technologies that has shown a lot of promise through its ability to capture information on surface protein levels.

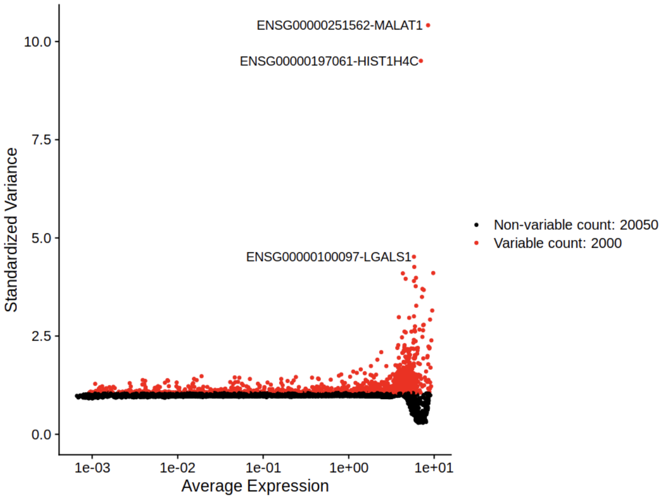


Fig. 22. Scatterplot for all genes in the dataset, examining their average expression versus their standardized variance. Each point represents a single gene, with variance being informed by all cells at all timepoints. Highly variable genes were selected for downstream timecourse analysis.

That being said, the introduction of this new level of protein information has not been reflected in the development of new visual encodings, at least in terms of popular toolkits such as Scanpy and Seurat. One of the main findings during the analysis process was that while there are some statistical methods that can be used to integrate gene expression and protein levels into a single dataset for analysis, the visualization idioms to convey the derived dataset are almost completely redundant the methods for gene expression analysis. As such it was decided to omit the inclusion of visualizations of the integrated protein information on the basis the interpretations would not change from gene expression

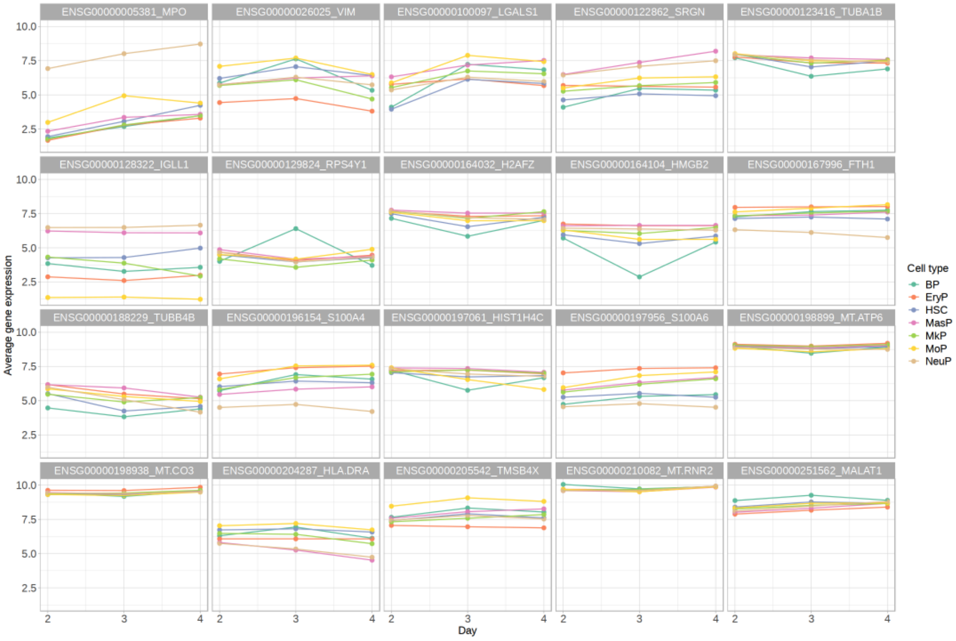


Fig. 23. Scatter plots for individual genes, plotting the mean expression for each cell type across all three timepoints. Data corresponding to each cell type is given a colour encoding. Each gene is given its own plot.

analysis alone.

Instead, a focus was placed on using visual encodings to investigate the protein levels over time, as that would provide insight to the relative changes in these functional units as cells differentiate. In particular, Fig. 24 depicts every single protein captured by the CITE-seq technology following aggregation of protein levels based on the cell type labels.

While most of the captured proteins were determined to have low expression levels, this visual encoding can capture interesting trends over time through the line mark. While most cell types had similar distributions of protein levels across the three days, there are some interesting trends that emerge with respect to the fluctuation of protein levels at particular time points (Fig. 24). This visual encoding is unable to determine whether the trends are due to noise with the sequencing technology, however it can capture a global overview of how protein levels within cells are in a constant fluctuation. The fluctuation captured by Fig. 24 may also represent variation within the cell type clusters, which could be observed through previous visualization idioms such as the DR plots.

6 Milestones

An overview of project milestones can be seen in Table 2.

7 Discussion and Future Work

This analysis project was able to demonstrate the broad application of visualization idioms that can be employed throughout the process of single-cell analysis. The number of research questions that can be investigated using these single-cell sequencing technologies is practically endless. Moreover, humans are unique biological entities in that there exists distinct patterns of variation within each persons genetic information. This variation likely manifests in different ways in different people at a hierarchy of levels, as cells work together to form organs, organ systems and eventually the human body. As such, the diversity in cells under investigation coupled with the vast number of research questions that can be investigated has made the analysis process convoluted and highly context-dependent. Seurat and Scanpy attempt to instill some degree of standardization to this analysis process, and there has been widespread adoption of these toolkits since their inception. The utilization of visualization to convey information has been deeply intertwined in the analysis process of cell sequences, and that role is likely to grow as more sequencing technologies that capture even more information are currently being developed.

As such, this project analyzes the different visual encodings that can be applied for the research question of evaluating cell type label clusters. A finding that was pervasive throughout the different visual encodings that were investigated was the lack of clear separation between the different cell types, both at the global and individual gene scales. Given that the dataset measures blood cells at the very early stages of differentiation, the lack of separation with respect to their molecular profiles was expected. The main motivation for assessing cells at such early time points is that the differentiation process that cells undergo as they mature to fill a specialized role is typically irreversible. So while clear cell type-specific patterns would emerge if a mature and differentiated cell population was sequenced, no insights can be derived with respect to the evolution of the biological system. Investigating the early stages of cell differentiation is particularly important in the context of disease, as identifying the exact developmental timing at which malicious cells begin to emerge can be used as a basis for developing therapeutic interventions to prevent such occurrences. Thus, while it is a challenging dataset to derive biological meaning, understanding the molecular underpinnings of cell differentiation over time offers much promise to basic biological knowledge and medical research.

We were able validate claims of in abstracted tasks, in that the different interpretations can be derived from different visual encodings. Importantly, those interpretations were also subjected to change over time as the cells differentiate into mature blood cells. The visualization idioms offered by Seurat, Scanpy and additional software packages can be applied to extract biological meaning from a number of different perspectives. The DR plots provide a global representation of expression levels of individual cells, which was conducive to evaluating cell type-specific trends at each time point. The interpretations derived from trajectory inference methods are much more aligned with the continuous nature gene expression across the cell population. While these global views remained qualitative, they provide important insights into the whether the assigned cell type labels were appropriate. In light of the similarity between different cell type labels in terms of their gene expression profiles, these methods that were able to capture the entirety of the high-dimensional dataset provided moderate support for the cell type labels provided.

Upon selecting a subset of genes to use as markers for identifying

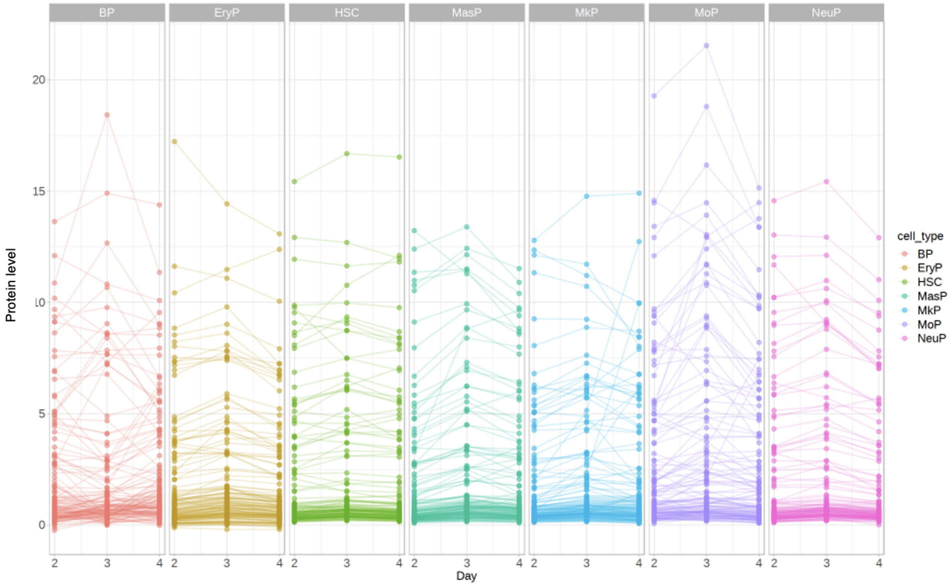


Fig. 24. Scatter plots for each cell type showing protein abundance across all three timepoints. Each cell type is given its own plot for all of the captured proteins. Cell type is visually encoded by colour.

Table 2. Milestones

Task	Estimated Time to Complete (hours)	Actual Time to Complete (hours)	Date Completed (YYMMDD)
Investigation into the Kaggle dataset and ideation around project questions	5	7	220925
Project proposal assignment	4	4	221018
Project update assignment	4	4	221113
Reading of the core papers relating to scanpy and Seurat	4	4	221014
Reading of the core papers relating to trajectory analysis	6	5	221023
Reading of the core papers relating to additional analysis such as DR-based graphical abstraction	6	6	221105
Reading of parallel papers performing related work	10	8	221110
Determine list of software tools that will be used and download them all any dependencies	10	6	221112
Data sanitation and preparation	4	4	221023
Seurat and scanpy analysis of dataset (including compute time)	30	45	221205
Trajectory inference analysis (including compute time)	15	15	221203
WaddingtonOT analysis (including compute time)	10	10	NA (too long to compute)
Graphical abstraction of UMAP/t-SNE (including compute time)	10	10	221011
Final presentation slides and recording	6	7	221213
Final paper writing	15	30	221215
Final paper figure generation	10	10	221215

the different cell types, it became apparent that there existed significant variation of gene expression within each cell type cluster. This conclusion was based on the finding that none of the marker genes selected by the methods tested could reliably discern every cell in one cell type from the rest of the cell population. However, the identified marker genes could be used to reliably capture subpopulations of cells within the assigned cell type labels, which suggests that further splitting the cells into more specific clusters may be warranted.

In terms of the visual encodings that standardized single-cell analysis packages like Scanpy and Seurat provide, there is a lack of support for visual encodings that capture trends in gene expression levels over time. There is a general trade-off for the visualization idioms that these single-cell analysis packages offer in that they can capture global-scale trends in gene expression profiles or very fine-grained analyses that assess a subset of genes. As such, this project identified a gap in visual encodings that could be addressed by future design projects. In devising an effective visual encoding that could fill this need, we determined that the current visualization idioms offered by Scanpy and Seurat do not provide satisfactory support for time-series analysis in an unbiased manner. Additionally, more visualization strategies could be developed to incorporate other levels of information (i.e., protein levels) captured for individual cells.

One potential idea for designing a new visual encoding could involve an initial step of clustering cells based on gene expression. Algorithms such as Leiden clustering [19], a generalization of K-means clustering called X-means clustering [13], or hierarchical density-based clustering [10] could be used to get an initial decomposition of the cell population. Ideally, the clustering algorithm would be unsupervised with respect to the number of clusters, and a higher granularity clustering algorithm would be preferred in order to capture as many cluster-specific trends as possible. Given these labels, we could then identify differentially expressed genes that can characterize each grouping, and these subsets of genes can then be aggregated and treated as a gene expression "modules" to represent each cluster, as in Velten et al. [20]. Genes that were not identified in the as part of any cluster-specific module can be aggregated as a "background" signal module. Together, these aggregated genetic modules can provide an lower-dimensional representation of the dataset that can be visualized with quantitative visual encodings such as trees, scatter plots, line plots and bar plots. To further prevent loss of information through aggregation, interactivity can be incorporated such that selection of a module can expand re-expand the subset of genes. Successful incorporation of such an approach would require extensive domain-specific knowledge and validation, but could be valuable in subjecting global, cluster-level trends to quantitative analysis.

8 Conclusions

Single cell RNA sequencing is a powerful tool in the cell biologists' toolbox, capable of deeply characterizing a given cellular system at the molecular level. Due to the incredible amounts of data produced, analysis and visualization is as much of a challenge as is collecting the data itself. Our body of work in this publication aimed to characterize the existing gold-standard tools for single cell analysis, Seurat and Scanpy, as well as packages aimed at more specialized analysis (WaddingtonOT, PAGA, and Monocle3). Throughout our exploration of our own analysis and others in parallel, a clear gap in the ability of these tools appeared. While these packages were ideal for either a high-level interrogation of the dataset, or a more zoomed in examination targeted at specific genes or cell types, there was a general lack of content bridging these two extremes. After discussing this gap, we proposed a potential visual encoding that potentially bridges this gap. While to date the field has been dominated by bigger and bigger datasets, we anticipate that as the field matures, more weight will be given to clever analysis and visualization, allowing researchers to get the most out of their hard-earned data.

Acknowledgments

We would like to thank Dr. Tamara Munzner and Mara Solen for all of the feedback and guidance throughout the term. In addition, we would like to thank Chenwei Zhang and Yibo Jiao (i-ViDa: Visualizing Energy Landscapes and Trajectories of DNA Reaction) for their time reviewing our project update and providing us with extremely helpful feedback.

References

[1] The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019. doi: 10.1038/s41586-019-0969-x

[2] A. Benz, A. Chow, D. Burkhardt, I. Jelic, J. Bloom, kvigly, lancerunstats, M. Luecken, P. Holderrieth, and R. Holbrook. Open problems - multimodal single-cell integration, 2022.

[3] C. Buccitelli and M. Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21:630–644, 10 2020. doi: 10.1038/s41576-020-0258-4

[4] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018. doi: 10.1038/nbt.4096

[5] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184:3573–3587.e29, 6 2021. doi: 10.1016/j.cell.2021.04.048

[6] B. Ho, A. Baryshnikova, and G. W. Brown. Unification of protein abundance datasets yields a quantitative saccharomyces cerevisiae proteome. *Cell Systems*, 6:192–205.e3, 2 2018. doi: 10.1016/j.cels.2017.12.004

[7] D. J. Knapp, C. A. Hammond, F. Wang, N. Aghaeepour, P. H. Miller, P. A. Beer, D. Pellacani, M. VanInsberghe, C. Hansen, S. C. Bendall, G. P. Nolan, and C. J. Eaves. A topological view of human CD341 cell state trajectories from integrated single-cell output and proteomic data. *Blood*, 133(9):927–939, 2019. doi: 10.1182/blood-2018-10-878025

[8] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne, 2008.

[9] E. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. Bialas, N. Kamitaki, E. Martersteck, J. Trombetta, D. Weitz, J. Sanes, A. Shalek, A. Regev, and S. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161:1202–1214, 5 2015. doi: 10.1016/j.cell.2015.05.002

[10] C. Malzer and M. Baum. A Hybrid Approach to Hierarchical Density-based Cluster Selection. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2020-September:223–228, 2020. doi: 10.1109/MFI49285.2020.9235263

[11] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. 2 2018.

[12] M. P. Mulè, A. J. Martins, and J. S. Tsang. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications*, 13:2099, 4 2022. doi: 10.1038/s41467-022-29356-8

[13] D. Pelleg and A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *CEUR Workshop Proceedings*, 1542:33–36, 2015.

[14] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015. doi: 10.1038/nbt.3192

[15] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(6):1517, 2019. doi: 10.1016/j.cell.2019.02.026

[16] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14:865–868, 9 2017. doi: 10.1038/nmeth.4380

[17] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, 2019. doi: 10.1016/j.cell.2019.05.031

[18] K. Takahashi and S. Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676, 2006. doi: 10.1016/j.cell.2006.07.024

[19] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 2019.

doi: 10.1038/s41598-019-41695-z

[20] L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz, E. C. Buss, D. Nowak, T. Boch, W.-K. Hofmann, A. D. Ho, W. Huber, A. Trumpp, M. A. G. Essers, and L. Steinmetz. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19:271–281, 4 2017. doi: 10.1038/ncb3493

[21] F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 12 2018. doi: 10.1186/s13059-017-1382-0

[22] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, 2019. doi: 10.1186/s13059-019-1663-x

[23] L. Yu, Y. Cao, J. Y. Yang, and P. Yang. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1), 2022. doi: 10.1186/s13059-022-02622-0