Assessing the Current, Future Landscape of scRNA-seq Data Exploration and Cell Type Modeling

By

KEITH MITCHELL

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

INTEGRATIVE GENETICS AND GENOMICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

David Segal, Chair

Megan Dennis

Ian Korf

Committee in Charge

2021

CHAPTER 1: The current state of single cell methods.

Scientific Breakthroughs in Single Cell Omic Analysis

Droplet-based microfluidic techniques in combination with continued sequencing advancement has revolutionized many areas of cell based research. In particular, microfluidics have allowed research in genomics to examine cellular structures such as developing embryos at a resolution previously unobtainable. The introduction of single-cell RNA sequencing (scRNA-seq) technologies and bioinformatics pipelines has enabled many breakthrough discoveries such as targets for immune system regulation, the Human Cell Atlas, and cell development trajectories and their associated gene expressions (Rozenblatt-Rosen et al. 2017; Reuell 2018). scRNA-seq is a rapidly growing approach for answering questions related to cell populations and their behaviors, responses, and peculiarities in important areas such as cancer, autoimmune deficiencies, and many other diseases. The answers that scRNA-seq can provide guide new therapeutics and the many new emerging technologies in single-cell genomics such as VDJ analysis and CITE-Seq have much more to offer. VDJ single-cell analysis targets the transcriptomic and immune receptor qualities and clonotypes of T and B cells. Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) allows for the quantification of proteins with the usage of DNA-barcoded antibodies (Stoeckius, Stoeckius, and Smibert 2017). These techniques along with many others are applied to bioinformatics processes and create additional data and necessary techniques for downstream analysis.

Importance of Applications for Data Visualization and Cell Typing

These many emerging technologies come with a variety of challenges in computational biology and bioinformatics. Data from single-cell analysis are often complex, containing many forms of metadata such as clonotype, cell types or subtypes, and one or more treatments. Translational bioinformatics is a crucial step for allowing specialized researchers to easily interact with the data and ask directed questions. In addition, tissue and cell samples can vary widely in cell type

contents, expression patterns, treatments and extraction. Though many methods exist for cell identification, these often are not one size fits all. In the following chapters two particular applications and methods will be discussed. The first is a method that uses generalized linear models and custom gene signature sets to identify cell types and shows improvement on existing methods. The second is an application that uses the R programming language package called Shiny that is popular for interactive data visualization. The application presented enables more complex visualizations than other available tools based on categorical and continuous variables in the dataset. Researchers can ask questions about gene expressions, treatments, clonotypes, cell types as well as their relative comparisons (for example the differences between a gene's expression in a cell type across various treatments). The specific views will be discussed in Chapter 3.

Difficulties in Cell Typing: CellularCall

As a brief overview, the cell assignment method has proven useful for complex samples. Benchmarking was performed on multinomial regression methods for cell classification in noisy scRNAseq data in order to compare Garnett and the proposed, improved method CellularCall (Pliner, Shendure, and Trapnell 2019). Garnett is a method that uses logistic regression based on prior knowledge, in this case marker genes and the relative cell types (Abdelaal et al. 2019). In particular, both methods were focused on scRNA-seq samples across multiple tissues, thus explaining the increase in noise, rare cell types, and numerous subtypes. A multinomial logistic regression, elastic-net regularization, and generalized linear model is used to train the classifier. With improvements in this method, a more consistent classification was obtained for CellularCall in regions where Garnett classified many unknown cell types with no particular pattern or methods for investigating the potential discrepancies in the model. Methods such as cell assignment help add useful metadata to data structures and objects.

Seurat and Tools For Visualizing Data: scRNA Shiny

A common data object used for single-cell methods is the Seurat object, which is an object from the popular R software from the Satija Lab at the New York Genome Center. The Seurat R toolkit for single cell genomics is one of the most popular packages for quality control, analysis, and exploration for many types of single-cell data. This is also the structure that the data visualization is built on since it creates a seamless transition from bioinformaticians to researchers. The information from bioinformaticians can then progress to researchers and specialists to investigate their research questions of interest. Along with Seurat, Shiny is a package in R that allows for construction of webpages, apps, and dashboards for interactive data visualization. This enables researchers to work with data in the browser which can be hosted by bioinformaticians and saves countless hours of software install and across operating system issues.

Current tools for scRNA-seq analysis graphic user interfaces are typically centered around point-and-click processing of data as opposed to dynamic exploration of processed data. Advanced views built around custom bioinformatics processing are becoming increasingly important with the continued increase in complexity of scRNA-seq experiments and their relative metadata of interest. Thus, scRNA Shiny is provided which provides many custom data representations to more thoroughly explore data provided with regards to the research questions. An Amazon Web Services instance is available for users to run their own shiny apps and enable easy access for others with a public link within their accounts. The app is also additionally easily run locally using a Docker container.

CHAPTER 2: Benchmarking multinomial regression methods for assignment of cell type to heterogeneous, noisy single-cell RNAseq data and an improvement on published methods, CellularCall (CC)

Keith Mitchell[1,2,3], Dustin Leale[2], Blythe Durbin-Johnson[1,3], Han Chen[1], Hans Mueller[1], Matt Settles

1. Division of Biostatistics, University of California, Davis, CA, USA
2. Department of Genetics and Genomics, University of California, Davis, CA, USA
3. Bioinformatics Core, Genome Center, University of California, Davis, CA, USA

ABSTRACT:

A plethora of tools exist for scRNA-seq cell assignment, yet these are often designed for single tissue samples and a predefined set of cells rather than two or more tissues and custom sets of expected cells. Multiple tissue samples and extraction techniques often contain heterogeneous and therefore noisy single cell RNAseq. This is due to a wider array of cell types expected and repeat cell types across multiple tissues and extraction. Some tools such as Garnett exist and allow for custom exploration of a custom set of expected cells and markers, yet such methods are often not flawless. Benchmarking was performed on multinomial regression methods for cell classification in noisy single-cell RNAseq from knee tissues in order to compare Garnett and a proposed, improved method CellularCall. In particular, these methods were focused on scRNA-seq samples across multiple tissues, thus explaining the increase in noise, rare cell types, and numerous subtypes. To be more specific a multinomial logistic regression, elastic-net regularization, and generalized linear model was used to train the classifier. A more consistent classification was obtained for CellularCall in regions where Garnett classified many unknown cell types with no particular pattern or methods for investigating the potential discrepancies in the model. CellularCall improved in performance and interpretation with regards to the k-nearest-neighbor clustering. Garnett unexpectedly had a large number of unknown calls randomly dispersed in Neutrophil clusters while CellularCall did not have these issues. Including more coefficients allowed the creation of a more complex model and proved to largely improve results. With regards to method improvements, CellularCall allows a user to select a value of α,

the mixing parameter, to best fit the complexity of their data. Further, probabilities of cluster? assignments are returned, which results in a much more informative analysis and diagnosis.

INTRODUCTION:

The introduction of scRNA-seq technologies and bioinformatics pipelines has enabled many breakthrough discoveries. scRNA-seq is a rapidly growing approach for answering questions about cell populations and their behaviors, responses, and peculiarities in important areas such as cancer, auto immune deficiencies, and many other diseases. Typically researchers attempt to identify and explore cells using marker genes based on clustering performed using the R package seurat and a shiny app or other explorative tool. This is a laborious process and other options exist for automatic cell type identification, therefore benchmarking tools for this purpose is crucial. The systematic benchmarking of omics computational and analytical tools has become crucial for helping bioinformaticians and biostatisticians provide the most proper and rigorously evaluated methods in biological research (Mangul et al. 2019). Previous work has been done for comparing a plethora of tools for cell identification of scRNAseq data, yet this evaluation was only considered for peripheral blood mononuclear cell datasets (PBMC). This sort of sample type is a simple, well studied cell grouping (Abdelaal et al. 2019). In addition, many of the methods considered rely on the assumption of well understood and simple tissue types as they are based on pre-trained neural networks (Davis n.d.) (Abdelaal et al. 2019). These neural networks supplied for individual types prove to not be robust in the face of novel, complex multiple-tissue datasets with varying extraction techniques and are difficult to extract important information from. Other methods such as SVM, LDA, KNN, and neural networks are either hard to interpret or do not translate well to cell groups that are rare like Chondrocytes.

There exists only one other multinomial classifier, Garnett, which allows for the ability to build custom models by training on a subset of a given dataset (Pliner, Shendure, and Trapnell 2019). It is worth noting that previous benchmarking has shown the performance is highly dependent on the marker set provided which is true in this benchmarking as well. Previous work has pointed out flaws in their approach and calling of “unknown” cell types which will be reiterated in this benchmarking; therefore, we present a refined multinomial regression method called CellularCall. This method was developed using a generalized linear model framework at the UC

Davis Bioinformatics Core with the hopes of creating a more robust solution for datasets of interest. The Haudenschild group (UC Davis) has a complex dataset composed of PBMC, bone matter, cartilage, ligaments and other cells from tissues found in the knee across multiple extraction techniques. The dataset is a matrix of 18,738 cells for the first dataset and 33,148 cells for the second dataset, which are identified by a unique barcode (rows) and are considered independent units of replication. In addition, columns or predictors are gene expression values of which there are ~17,000 depending on filtering strategies.

In this paper, CellularCall is performed in a comparison to Garnett in regards to the Haudenschild knee dataset. The comparison of the two methods is for the purpose of showing the need for flexibility in the elastic net regression parameters. By varying the balance between ridge and lasso regression we aim to show the improved assignment of cells for this particular dataset. In addition, there is a need for more informative results such as probability of given cell assignments. This will aid researchers when analyzing significant coefficients, inconsistencies, and other complexities such as areas of high noise in the model.

METHODS:

The data was normalized using log normalization, a common practice in scRNAseq, prior to comparison of methods of interest (Hafemeister and Satija 2019; "Seurat Part 3 – Data Normalization and PCA" 2018); in addition, a set of marker genes will be used in a multinomial logistic regression model to predict the cell type of the cells in the data set.

Multinomial Classification

To begin, a multinomial logistic regression, elastic-net regularization, generalized linear model is used to train the classifier. $X_i = (x_{i1}, x_{i2}, , x_{ip})$ with $i = 1, 2, \cdots, n$ where n represents the number of cells in our samples and p represents the number of genes. $y_i \in \{1, 2, \cdots, M\}$ where k represents the number of possible cell classifications or cell types (as well as the possibility for unknown classification). $\beta_m = [\beta_{0,m} + \cdots + \beta_{p,m}]$ where β_p, m is the coefficient for the pth explanatory variable and the mth outcome that represents the log odds

ratio and $1 \leq m \leq M$ while $2 \leq j \leq M$ (Wikipedia contributors 2020). Consider $\pi_{ij} = Pr(Y_i = j)$.

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = 1)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i = \eta_{i2} = \ln \frac{\pi_{i2}}{\pi_{i1}},$$

$$\cdots, \ln \frac{\Pr(Y_i = M)}{\Pr(Y_i = 1)} = \boldsymbol{\beta}_M \cdot \mathbf{X}_i = \eta_{iM} = \ln \frac{\pi_{iM}}{\pi_{i1}},$$

$$1 = \sum_{j=1}^{M} \pi_{ij}$$

$$\pi_{i1} = 1 - \sum_{j=2}^{M} \pi_{ij} = 1 - \sum_{j=2}^{M} \pi_{ij} \cdot exp(\boldsymbol{\beta}_M \cdot \mathbf{X}_i)$$

Where 1 is the first possible classification and represents the baseline. The entire proof is not shown here for purposes of brevity. The baseline was arbitrarily chosen to be Basophils while benchmarking. This leads us to the following equation:

$$\pi_{i1} = \frac{1}{1 + exp(\boldsymbol{\beta}_M \cdot \mathbf{X}_i)}$$

$$\pi_{i1} = \frac{1}{1 + \sum_{j=2}^{M} exp(\eta_{ij})}, \pi_{ij} = \frac{exp(\eta_{ij})}{1 + \sum_{j=2}^{M} exp(\eta_{ij})}$$

Generating Coefficients

Maximum *a posteriori* estimation (MAP) is closely related to maximum likelihood with a regularization technique applied. Generalized iterative scaling, coordinate descent algorithms, and iteratively reweighted least squares (IRLS) are just a few of the methods used to approximate the coefficients.

Elastic net regularization for high dimensional data as opposed to LASSO and Ridge method (selecting a proper α)

The elastic net method is used which combines (least absolute shrinkage and selection operator) LASSO and ridge methods in order to create a linear combination of the two for the penalty function. This method allows for a mix between LASSO, which works best with few significant

predictors expected whereas ridge works best with all predictors expected to be significant. Garnett only allows for elastic net regularization ($\alpha = 0.3$) but this returns a limited number of coefficients (~45) which do not nearly encapsulate the complexity of the data and different cell types; therefore, CellularCall utilizes $\alpha = 0.1$.

$$\hat{\beta} = argmin(\sum_{i=1}^{n}(y - X\beta)^2 + \lambda \left(\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1 \right))$$

K-nearest neighbor (KNN) clustering based on principal component analysis (PCA) selection as a benchmark for cell assignment

Seurat uses a graph-based clustering technique based on principle components in order to apply K-nearest neighbor clustering means clustering at various resolutions. Resolution refers to the "granularity of the downstream clustering, with increased values leading to a greater number of clusters". This has become a common technique for scRNA-seq data (Xu and Su 2015)(Macosko et al. 2015). The number of PCAs used is determined by the user in order to determine a euclidean upon which the KNN algorithm is conducted. PCAs used for the Haudenschild dataset were determined using a scree plot, of which the first 29 were chosen. This was performed using the Seurat package for R (Hafemeister and Satija 2019). A range of resolutions is chosen and therefore there is a range of cluster numbers to be assessed. This method will serve as a general reference for the performance for the GLM, though KNN often fails to separate out rare cell types.

Uniform Manifold Approximation and Projection (UMAP)

The UMAP plot has become a favorite for representing scRNAseq data and is similar to another unsupervised machine learning algorithm for representation of high dimensional data called the t-distributed stochastic neighbor embedding (t-SNE) (McInnes, Healy, and Melville 2018; Wikipedia contributors 2021). UMAP was created using the Seurat package for R (Hafemeister and Satija 2019). UMAP and t-SNE are used to represent high dimensional data in two dimensions. This is purely a visualization technique and the core of the clustering from Seurat is performed using KNN clustering based on PCA, as mentioned in the previous section.

RESULTS:

Previous analysis showed that Garnett's multinomial regression performance over-classified the number of "unknown" cell types in both the original dataset from the Haudenschild group (dataset 1) as well as the new dataset (dataset 2, Figure 1 and Figure 2). Neutrophils were identified using marker genes, clustering, and manual assignment of clear cell types. There appears to be no pattern to the unknown calls within groups like Neutrophils and the classification of general "unknown" is not helpful with informing how to improve the model at this stage of classification and analysis.

CellularCall was applied to the new dataset supplied by the Haudenschild group, following the preprocessing. This yielded the results shown in Figure 3 and Figure 4, where rather than classifying cells as "unknown" during training and classification this proposed method suggests overlapping probability of cell calls on the UMAP for further diagnosis. The right panel in Figure 4 shows the probability for the calls in the left panel; although additionally, users could easily access this probability-based graph on a per cell-type basis to help diagnose where there may be unwanted overlap between cell types in the model. Figure 5 and 6 also shows the CellularCall performance and clustering split by the original sample identity of which there are 4. Overall, cell classifications seem consistent across tissue digests and knees and certain cell types are altered or only observed once across the two digests. Lambda is varied, evaluated, and optimized using cross validation in the glmnet() package as part of training the logistic regression (Figures 7 and 8).

DISCUSSION

CellularCall performance for most cell types are consistent across the four samples despite large differences across the tissue digests when displayed in the UMAP, for example the different groups of neutrophil cells observed. The value of the mixing parameter, α, was decreased. This creates an elastic net regression that has closer resemblance to ridge regression as opposed to lasso regression. Decreasing the value of α and including more coefficients to create a more complex model proved to have a large improvement on the results. Implementation of an "unknown" group should take place on the probability of assignment after the classification using the model is performed as the "unknown" is uninformative for diagnosing the model and

the results. Baseline options as well as values for alpha and lambda should be more fully considered by the user. In fact, one might argue it is incorrect to implement an “unknown” type during the training stage as there could be multiple cell types existing as “unknown” type. CellularCall improved in performance and interpretation with regards to the KNN clustering in addition to probability of calls rather than the hard to interpret and likely hard to model “unknown” cell type. Garnett has a large number of unknown calls randomly dispersed in Neutrophil clusters which is unexpected, but CellularCall did not have these issues.

FUTURE DIRECTIONS

Adjusting the alpha from 0.3 to 0.1 proved to be successful (for this dataset) therefore this should be an option to users in the future. This feature will naturally occur as further effort is made to make the workflow presented here into functions with more official documentation. Similarly, the baseline category should be further explored and allowed to be varied by a user to check performance. With regards to benchmarking, more tissue types should be tested individually, such as simple well studied tissue types. Simulated datasets should be composed of prelabeled datasets across multiple tissues with an array of noise for testing. Datasets evaluated in this study results in UMAP with a large spread compared with others, so varying degrees of this should be tested. Evaluation is applied to more complex datasets and benchmarking scope should be expanded to include DigitalCellSorter and SCINA. These are, respectively, a voting based on cell type markers and a bimodal distribution fitting for marker genes (Abdelaal et al. 2019). Further consideration should be made in the benchmarking practices for pre labeled simple tissue types and multiple simple tissue types put together.

APPENDIX:

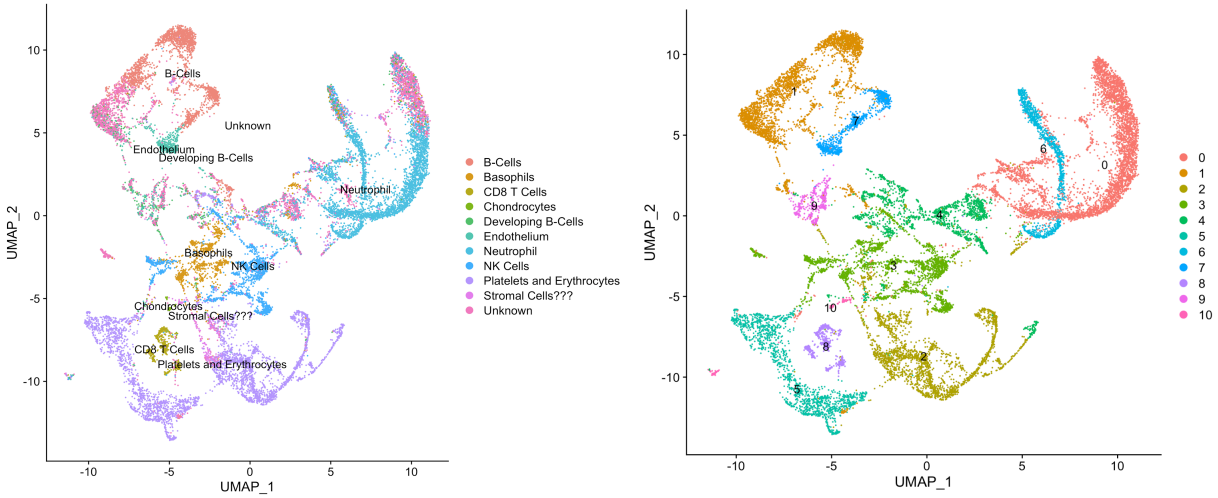


Figure 1: Revisiting the analysis performed on the previous dataset. The left panel shows the Garnett cell type classifications. Representation of a KNN PCA based clustering where k=11 and generally matches the number of cells we would like to apply the multinomial classification to when including the unknown type as an option. The UMAP multidimensionality reduction visualization is used here to aid with interpretation.

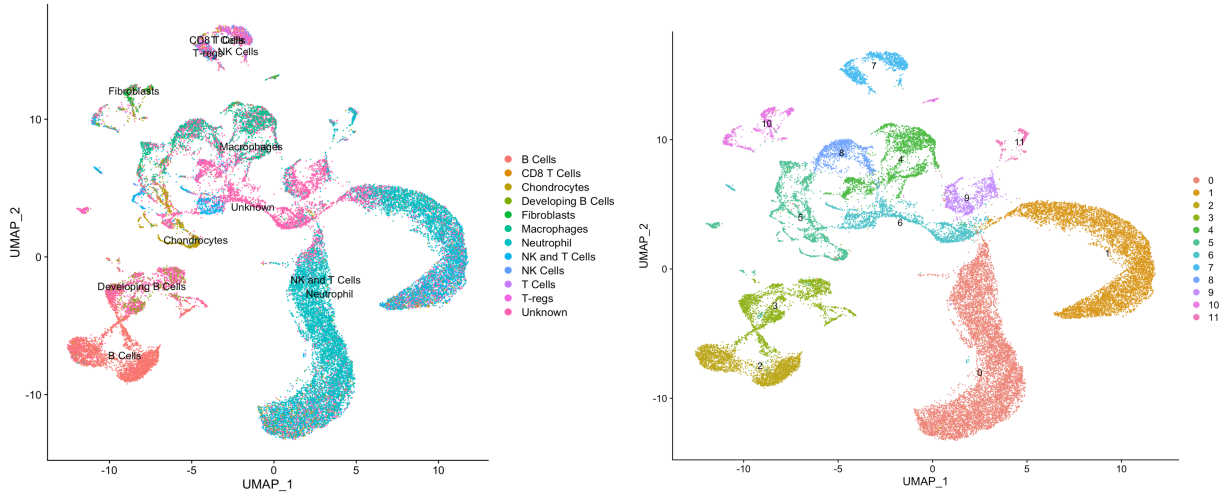


Figure 2: Garnett when applied to the new dataset from the Haudenschild group. An overlay of the cell type classifications on the UMAP multidimensionality reduction visualization. The right panel shows the KNN PCA based clustering for k=12.

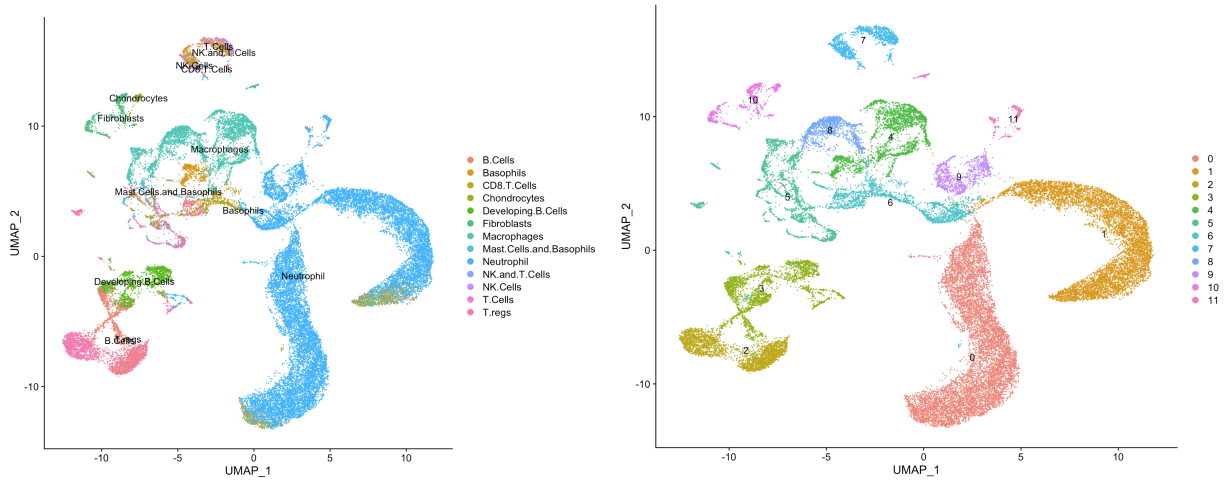


Figure 3: An overlay of the cell type classifications from CellularCall on the UMAP multidimensionality reduction visualization when applied to the new dataset from the Haudenschild group. The right panel shows the KNN PCA based clustering for k=12 for comparison.

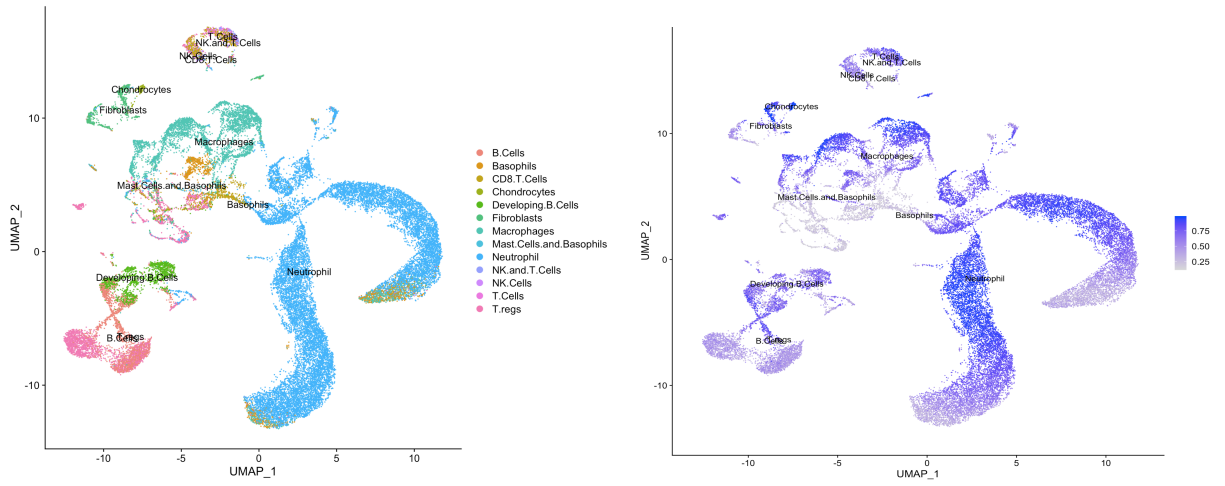


Figure 4: An overlay of the cell type classifications from CellularCall on the UMAP multidimensionality reduction visualization when applied to the new dataset from the Haudenschild group. The right panel shows the probability of each call on a scale of 0-1 on a UMAP.

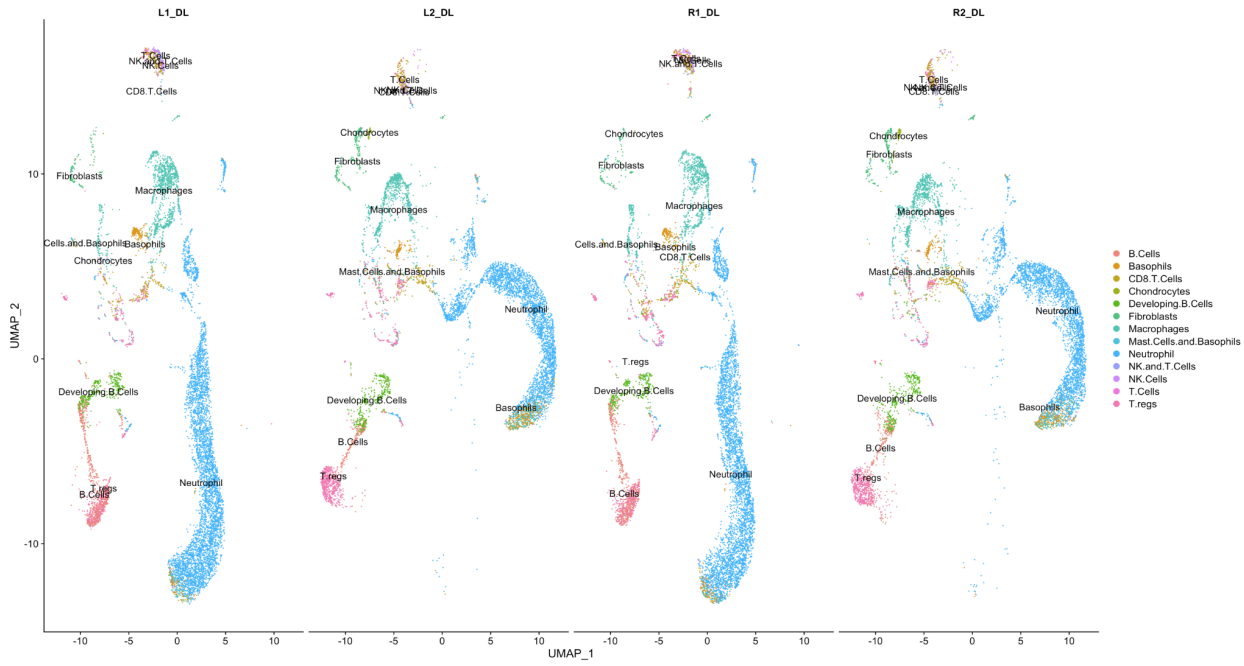


Figure 5: CellularCall cell type classifications across the four samples included in the new dataset from the Haudenschild group. L indicates that the same was taken from the left knee and R indicates the sample was taken from the right knee. 1 indicates the samples are from the first tissue digest while 2 indicates the samples were from the second tissue digest.

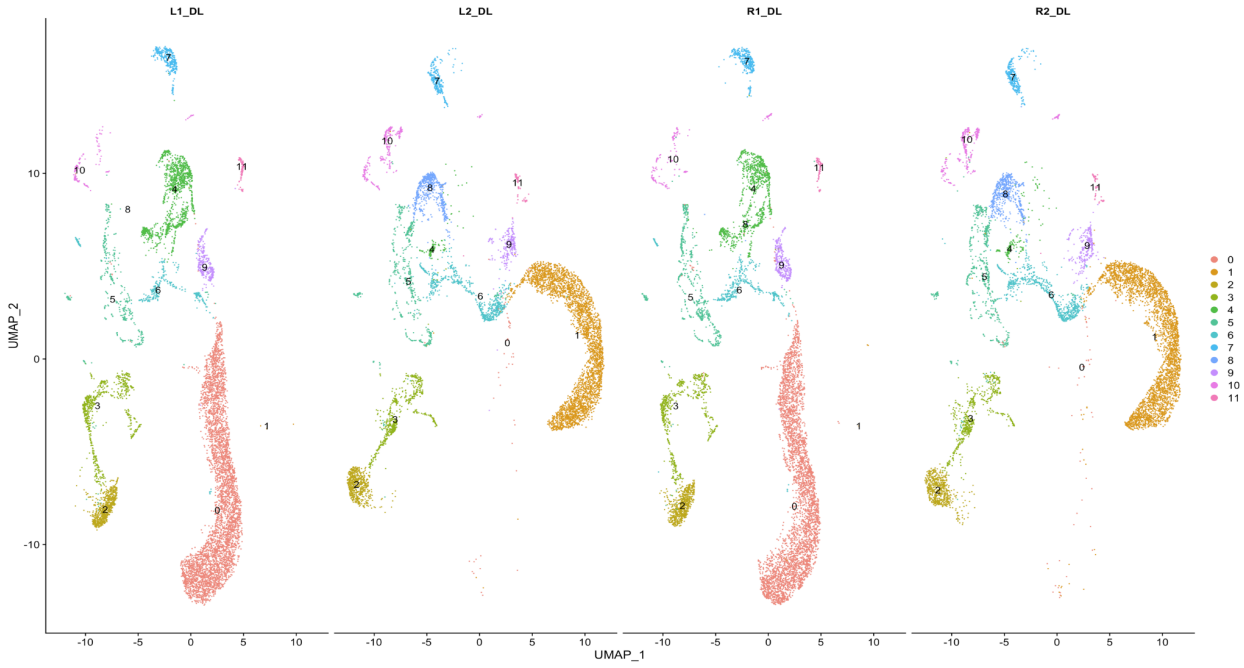


Figure 6: KNN PCA-based clustering with k=12 across the four samples included in the new dataset from the Haudenschild group. L indicates that the same was taken from the left knee and R indicates the sample was taken from the right knee. 1 indicates the samples are from the first tissue digest while 2 indicates the samples were from the second tissue digest.

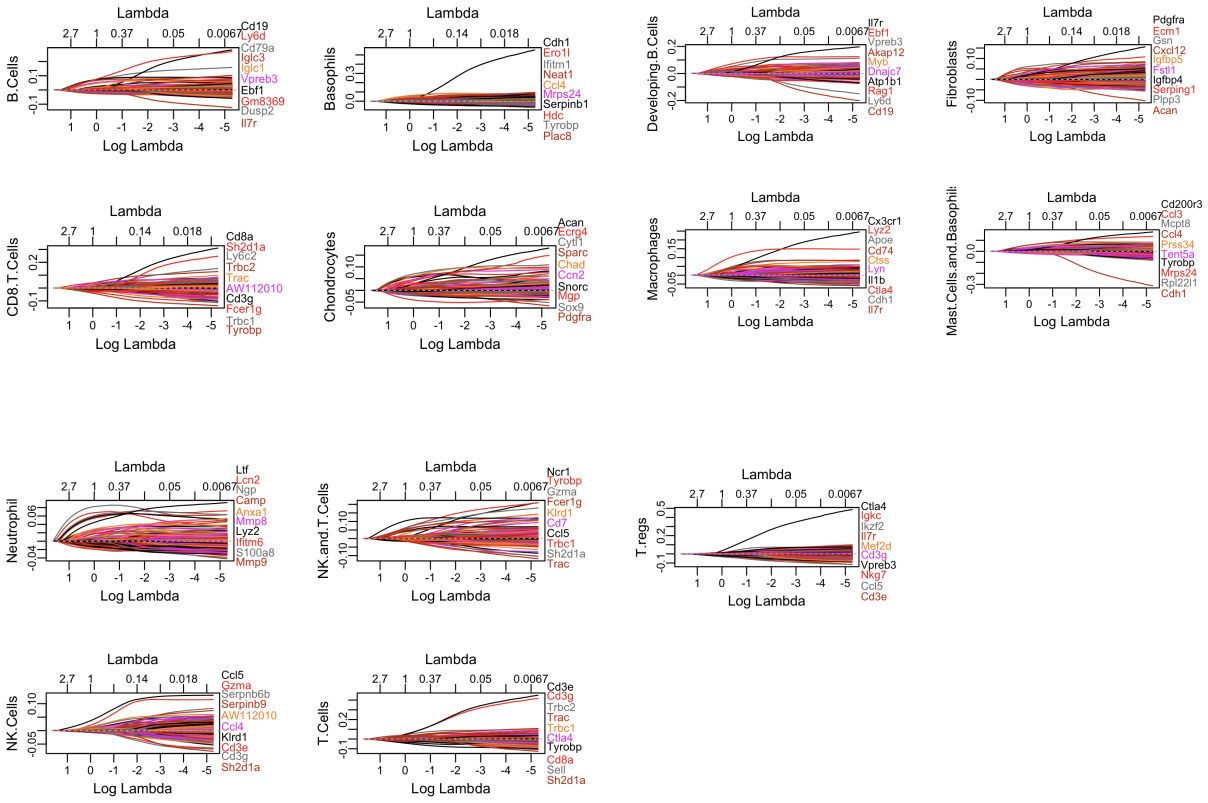


Figure 7: Representation of coefficient values for given genes based on the $log(\lambda)$ values.

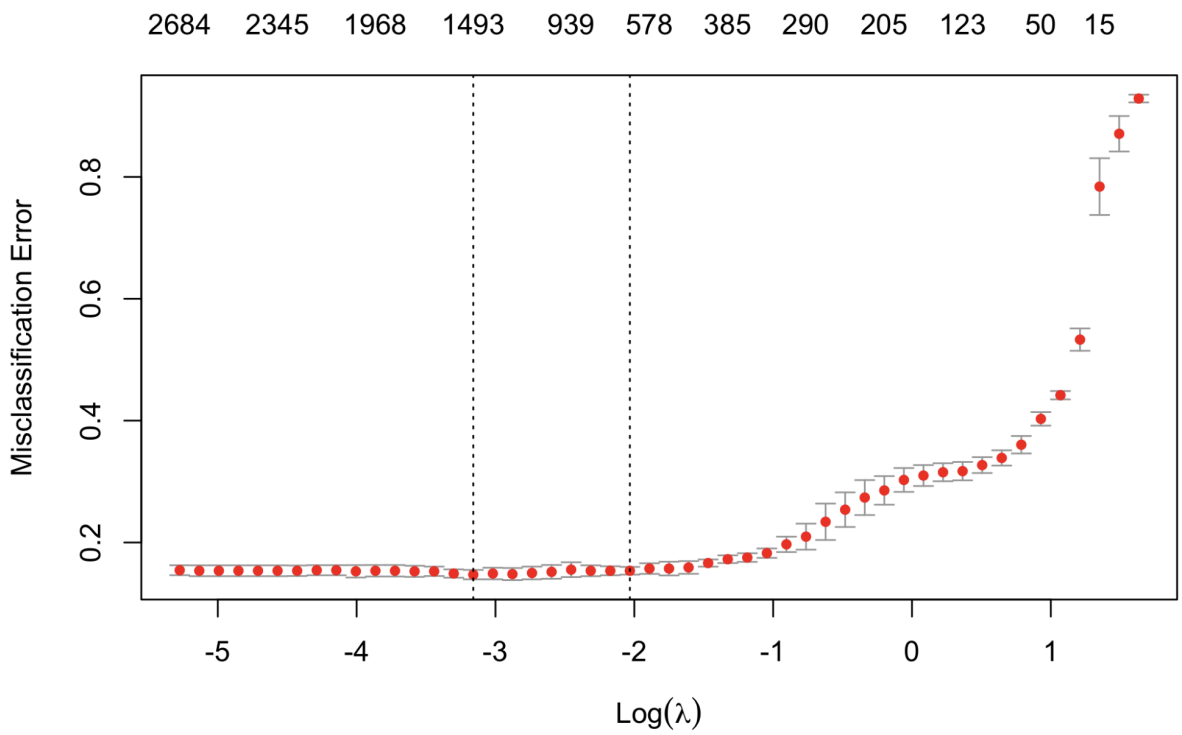


Figure 8: Plot of cross validation errors against candidate regularization parameters for an elastic net logistic regression. The first vertical line is λ_{min} while the second vertical line is λ_{1se}. The numbers on the top y-axis are the number of nonzero coefficient estimates for a given value of $log(\lambda)$.

SUPPLEMENTARY MATERIALS:

Github link: https://github.com/keithgmitchell/CellularCall

CHAPTER 3: scRNA Shiny: bridging advanced data analysis and data exploration of scRNA-seq projects between bioinformaticians and biologists

Keith Mitchell, Blythe Durbin-Johnson, Dustin Leale, Monica Britton, Sam Hunter, Matt Settles

1. Division of Biostatistics, University of California, Davis, CA, USA
2. Department of Genetics and Genomics, University of California, Davis, CA, USA
3. Bioinformatics Core, Genome Center, University of California, Davis, CA, USA

ABSTRACT

Current graphic user interfaces for scRNA-seq analysis are typically centered around point-and-click processing of data as well as creation of simple visualizations of this form of data. Here, we present scRNA Shiny with the purpose of more advanced exploration of processed data as compared to other available tools. Advanced views built around custom bioinformatics processing are becoming increasingly important with the continued increase in complexity of scRNA-seq experiments. Thus, scRNA Shiny is presented which provides many custom data representations to more thoroughly explore data provided with regards to the research questions. An Amazon Web Services instance is available for users to run their own shiny apps and enable easy access for others with a public link within their accounts. The app is also additionally easily run locally using a Docker container.

INTRODUCTION

scRNA-seq is an increasingly popular technology for answering questions about tissues, the cells present, and their relative expression. As technology usage increases, complexity of the usage is also seen. Some examples of this seen in scRNA-seq are VDJ analysis (which might take the form of clonotypes), expression analysis of treatments (i.e., KO vs WT experiments), software based cell-assignment methods, and transcript level analysis along with expression. Many of these options require working with the metadata of the experiment using complex comparisons

and visualizations, but not many options exist for exploring this sort of data following these techniques.

The importance of efficiency in the link between bioinformatics and biologists for single cell experiments is rapidly growing. In addition, with the advancement of single cell technologies there is an increase in complexity of research questions. Many bioinformaticians lack the subject-matter knowledge to explore the data with a nuanced perspective of all biological questions to be considered. In contrast, many biologists lack the ability to manipulate objects and install software for ease of data exploration. Therefore, flexible linkages between these two kinds of researchers are missing in the context of scRNA-seq experiments.

A plethora of tools currently exist for visualization and processing of Seurat objects. For example, one such popular tool is NASQAR's Single-Cell RNA app called SeuratV3 Wizard. Though this tool is comprehensive, it lacks the ability for unique and custom usage of Seurat, project metadata, or cell cycle control as part of processing. In addition, it does not offer many views that are often helpful for generating the variety or quantity of preliminary plots needed for exploring advanced research questions (Yousif et al. 2020). Due to the limitations and lack of versatility of web based platforms for processing workflows as well as the potential large size of datasets, it is not practical to run this processing on anything but high performance computing clusters (HPC) with scripts which can be easily altered by a bioinformatician.

scRNA Shiny was created using Seurat, Shiny, and the Shiny web server (Hao et al. 2020) ("Shiny" n.d., "Shiny Server" n.d.). The application was developed over a series of experiments for researchers where views for the application were organically created in order to resolve necessary forms of exploration. Questions of interest for the researchers can be answered via these views while saving time for bioinformaticians as well as the researchers themselves. The app is made to be intuitive while providing a suite of views to explore relevant portions of the data.

METHODS

Significant terminology used in the app and that helped guide development methodologies include *marker*, *feature*, *gene*, *numeric metadata, resolution*, *identity*, *reductions*, *t-SNE*, *UMAP*, and *PCA*…

Marker typically refers to genes in the data that are representative of a cell type of interest in the sample. Markers or marker genes are subsets of the numeric data. In general, a good marker gene for some identity should have a high mean expression and a high percentage of cells expressing the gene for a cluster of some identity chosen. In addition, said marker should result in a low mean expression as well as a low percentage of cells expressing the genes for the rest of the cluster for some given identity.

A *feature* of the data refers to some sort of numeric factor of the data and can be the expression of some gene or more generally metadata of the Seurat object of the numerical type. For example, the term feature may refer to the number of genes per cell for each sample or number of unique molecular identifiers per sample.

In general, the term *resolution* refers to clustering that was performed by Seurat where some particular clustering at some resolution is considered an identity. *Identity* is often certain clustering resolutions performed by Seurat or other categorical metadata such as cell cycle or automatic cell classification using tools such as Garnett (Pliner, Shendure, and Trapnell 2019). Seurat uses a graph-based clustering technique in order to use K-nearest neighbor clustering means clustering at various resolutions, which has become a common technique for scRNA-seq (single cell RNA-seq) data (Xu and Su 2015)(Macosko et al. 2015). The number of PCAs used of KNN is determined by the user in order to determine a euclidean upon which the KNN algorithm is conducted.

Reductions are simplified projections of the data to serve as a visual aid and representation of high dimensional data. The UMAP plot has become a favorite for representing scRNAseq data and is very similar to another unsupervised machine learning algorithm for representation of high dimensional data called the t-distributed stochastic neighbor embedding (t-SNE) (McInnes,

Healy, and Melville 2018; Wikipedia contributors 2021). *UMAP* and *t-SNE* are used to represent high dimensional data in two dimensions. This is purely a visualization technique and the core of the clustering from Seurat is performed using KNN clustering based on *PCA*, as mentioned in the previous section.

An Amazon Web Services (AWS) instance snapshot with a pre-configured configuration for the Shiny web server is available upon request for others to import to their accounts. This will allow users to create shareable links to quickly collaborate and explore data, and can easily be adjusted based on the size of the data in question. In addition, a docker container is provided so that local versions or other types of web services can be used to run the app with a few simple commands without across operating system installation issues.

By creating a set of terminology which all possible data fits into such as *marker*, *feature*, *gene*, *numeric metadata, resolution*, *identity*, *reductions*, *t-SNE*, *UMAP*, and *PCA* sets of views were explored. This happened organically over iterations of hypothesis exploration with researchers. It was necessary to create this simplification of data types and terminology in order to create sets of views that would work across many single cell studies.

RESULTS

The resulting application interface is composed of nine views including Documentation, Double Marker View, Single Marker View, Marker Set (Grid) View, Multiple Feature Plot View, Cluster Tree View, Separated Feature View, Separated Categorical View, and Finally the Marker Table View. (Figures 1-7) The dataset presented here is from a study with focus on exploring the immune system in four healthy canines with the goal of eventually comparing this data with the immune system in tumors. scRNA-seq was performed on these samples in addition to VDJ analysis. Some example resulting expression data and metadata is explored in the views presented. Alpha diversity of clonotypes at various clusterings was obtained and is easily explored using a variety of the views that support a continuous or “numeric” variable. In addition, researchers were interested in the average TRA and TRB sequence length as well as the

average number of TRA and TRB sequences at various clusterings of interest. This dataset demonstrates the uniqueness, complexity of the data and research topic.

The first and most basic result produced by the app is the Single Marker View (Figure 1). The first plot is a tSNE/PCA/UMAP which is chosen based on the reduction that is chosen, and is colored based on the Primary Numeric selection. The second plot is a violin plot that displays the Identity selection on the X-axis and the Primary Numeric on the Y-axis. The third plot is the tSNE/PCA/UMAP that is colored based on the Identity selection.

This is followed by the creation of a Double Marker View (Figure 2). All of the options here are the same as the Single Marker View but with secondary numeric as an option. Secondary Numeric, in combination with the Primary Numeric field, enables a user to explore two Genes, Numeric Metadata, or PCs based on the value selected for 'Numeric Analysis Type'.

In order to explore sets of markers, the Markers Set view was created (Figure 3). Y-axis represents the Identity, such as the original samples, cell assignments, or some groupings at a certain resolution. X-axis represents the genes selected (Primary Numeric). The size of each dot on the grid represents the percentage of cells that expressed that gene. The color intensity of each dot on the grid represents the average expression of the cells that expressed a given gene. This was followed by the creation of the Multiple Feature Plot (Figure 4). With this view, a CSV list of genes can be provided to create plots where >5 genes and ≤ 16 is optimal. The first plot is an identity of interest and the remaining plots represent the genes/markers of interest for exploration. With these two views it is possible to quickly summarize markers on top of a UMAP as well as get a clear statistical representation of these identities for a given feature.

Often when manual clustering is performed at various resolutions, it is necessary to visualize how these clusters change at various resolutions to aid with picking the optimal one. For this purpose the Cluster Tree view was created (Figure 5). This plot helps to identify closest related clusters so when moving into the final analysis you have a better idea of what the real cell groups are in your samples. However, note that in the Seurat workflow clustering is not performed on

TSNE or UMAP coordinates, but rather on principal components, with TSNE or UMAP used for display.

Separated Feature is a per identity panel view where a numeric data type of the data can be explored (Figure 6). Reduction is chosen to be UMAP in this case and identity is some clustering at some resolution provided by seurat. The primary numeric analysis in this case is the gene *CD8A* and allows for users to see samples separated based on their original identity. Often individuals will want to see how certain features or genes of their data behave across the samples taken. The first row of plots is the gene or numeric metadata feature highlighted across each of the samples. The next row is the clustering or identity of interest across each of the samples. Finally, the table at the bottom represents an identity versus samples of interest with regards to average expressions and percentage of cells expressed.

Separated Categorical is a per identity panel to test the interactiveness of these identities with another categorical variable (Figure 7,8). Reduction is chosen to be UMAP in this case and identity is some clustering at some resolution provided by seurat. Often individuals will want to see how certain identities interact with other identities in the data set. The first row of plots is the first identity highlighted across each of the samples. The next row is the secondary identity of interest across each of the samples. Finally, the heatmap at the bottom represents an identity versus identity heatmap where the frequency is shown for each of the possible combinations of the two identities. This allows, for example, a user to quickly explore which cell types are more highly abundant in certain samples or within groups of samples with varying treatments.

Now that the views available for scRNA Shiny have been explored, we see that the purpose and capabilities far exceed those available with NASQAR's SeruatV3 Wizard (Figure 9). SeuartV3 Wizard tends to lean towards the primary purpose of processing, filtering, and performing data reduction on small datasets with no ability to add custom metadata throughout the processing. The data exploration across various features of the metadata, such as identities, genes, categorical metadata, and numeric metadata. In comparison, scRNA Shiny is designed to be a post processing data exploration with much more capabilities than that seen in Figure 9.

There exist quite a few other apps for scRNA-seq data exploration. One very popular software that is extremely reactive and well written is CellxGene. CellxGene requires scanpy, and is not as malleable for across sample data exploration such as those presented in the Separated Feature and Separated Categorical View (Figure 10) (*Cellxgene: An Interactive Explorer for Single-Cell Transcriptomics Data* n.d.). Another existing tool, CellView, is shown in Figure 11. The hosted app is very limited for the size of the data that can be uploaded and additionally the tools is limited to similar views to the Single Marker and Marker Set views that scRNA Shiny has to offer (Bolisetty, Stitzel, and Robson 2017).

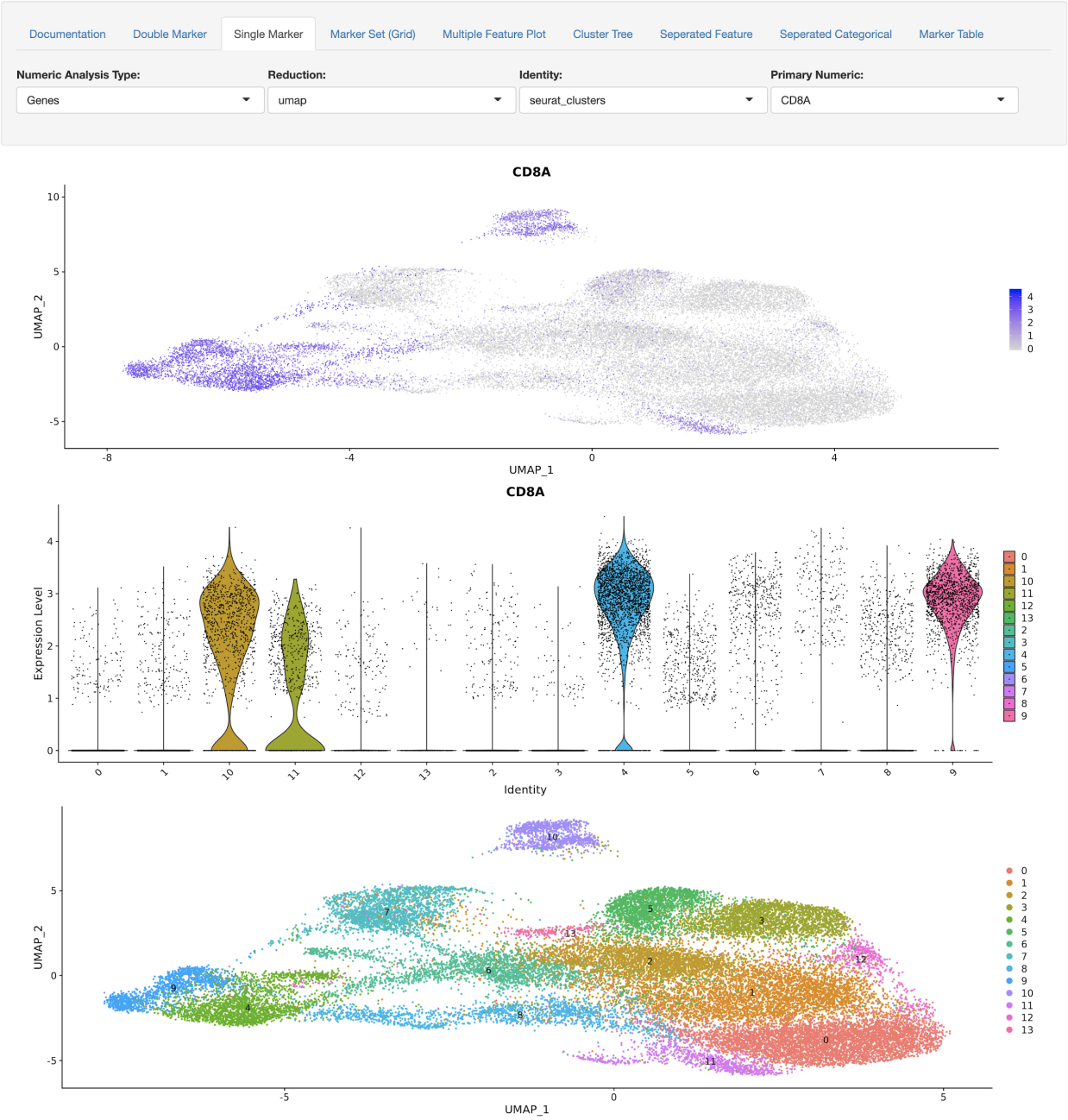


Figure 1: Single Marker View. Explore a single feature or numeric analysis type (gene, metadata, etc.) and its relation to variations of clustering or on a per sample basis.

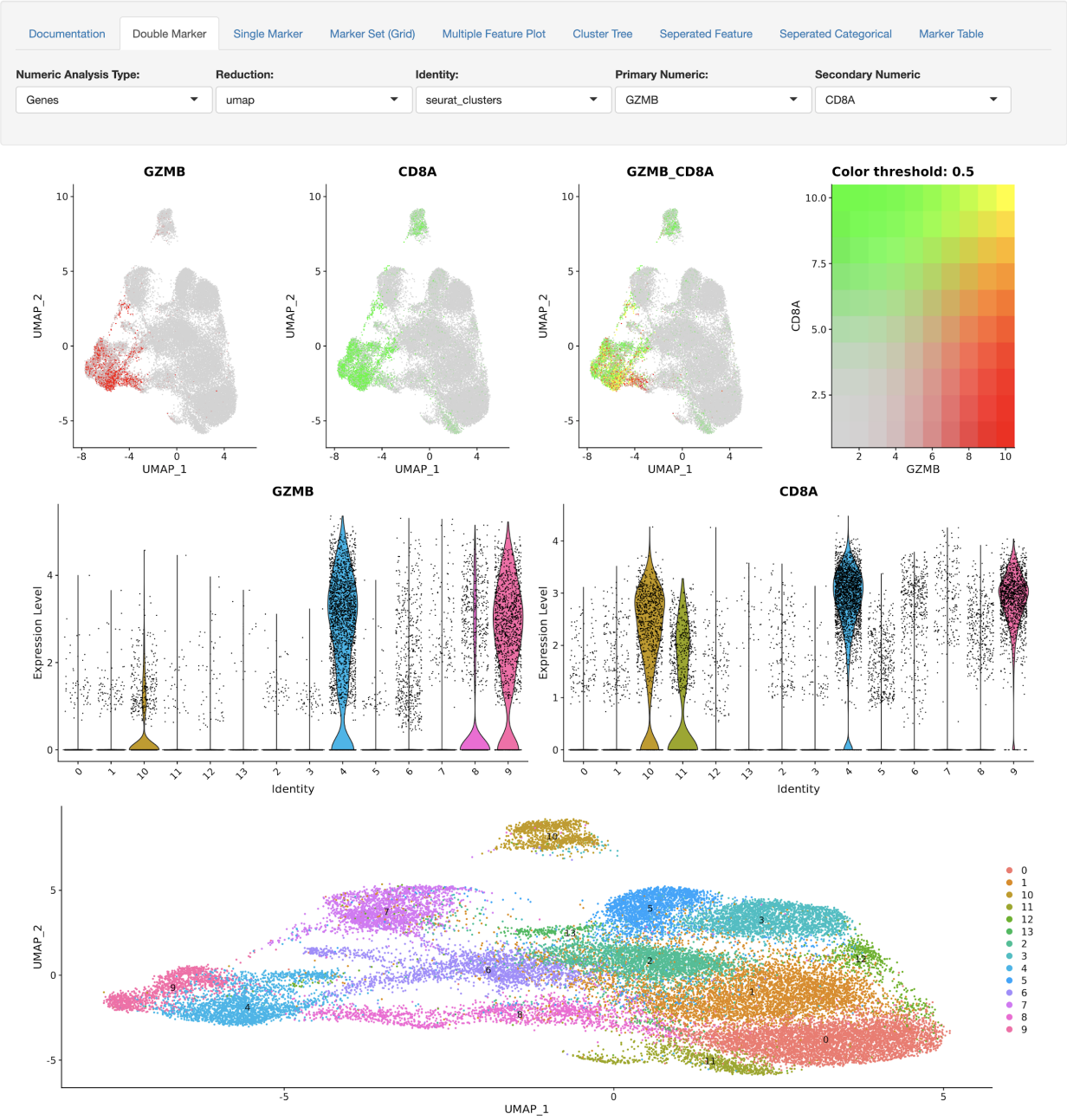


Figure 2: Double Marker View. Explore two features (gene, metadata, etc.) and its relation to variations of clustering or on a per sample basis.

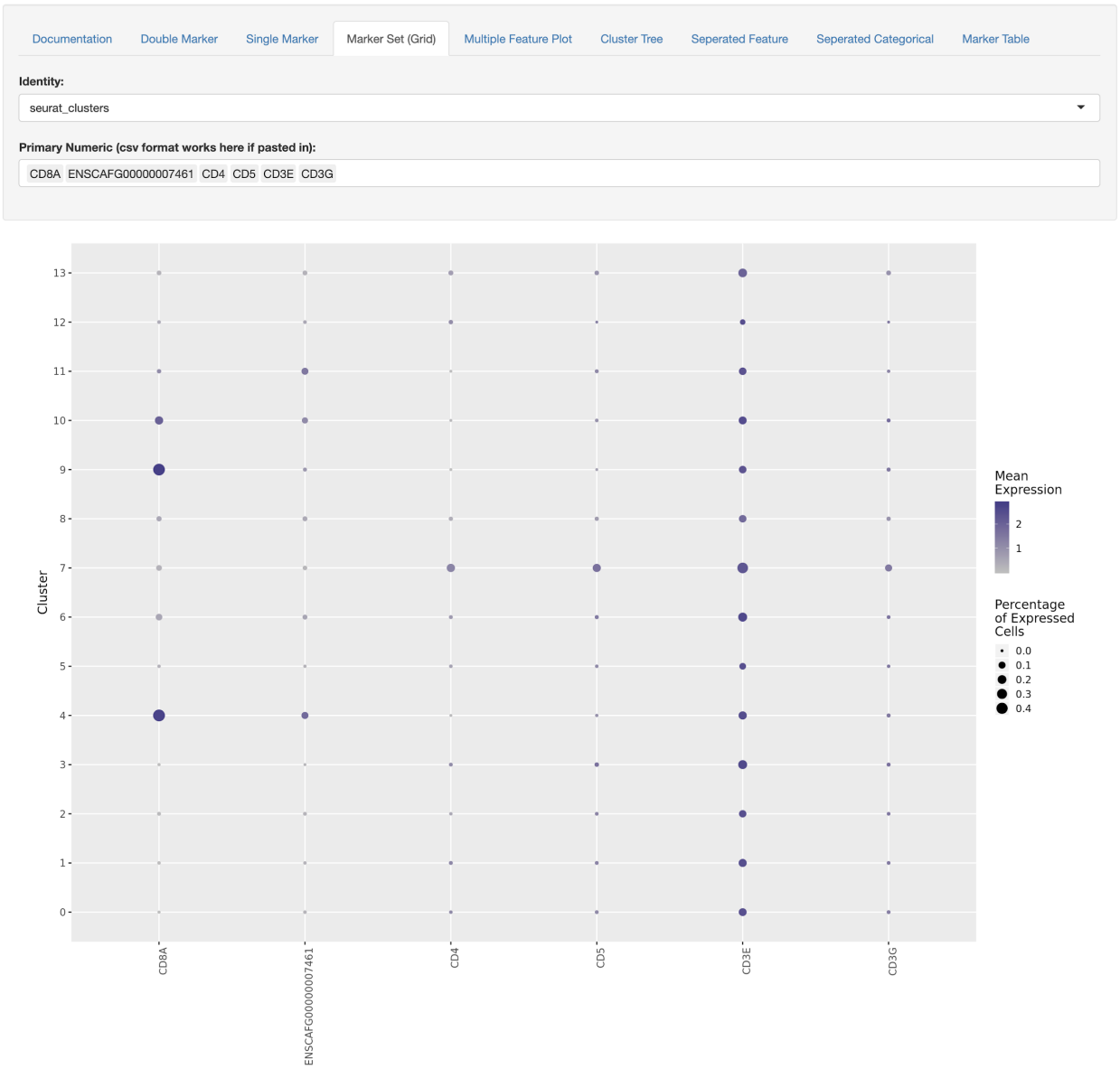


Figure 3: Marker Set, exploring sets of marker genes and their relation to an identity, or clustering/cell assignment.

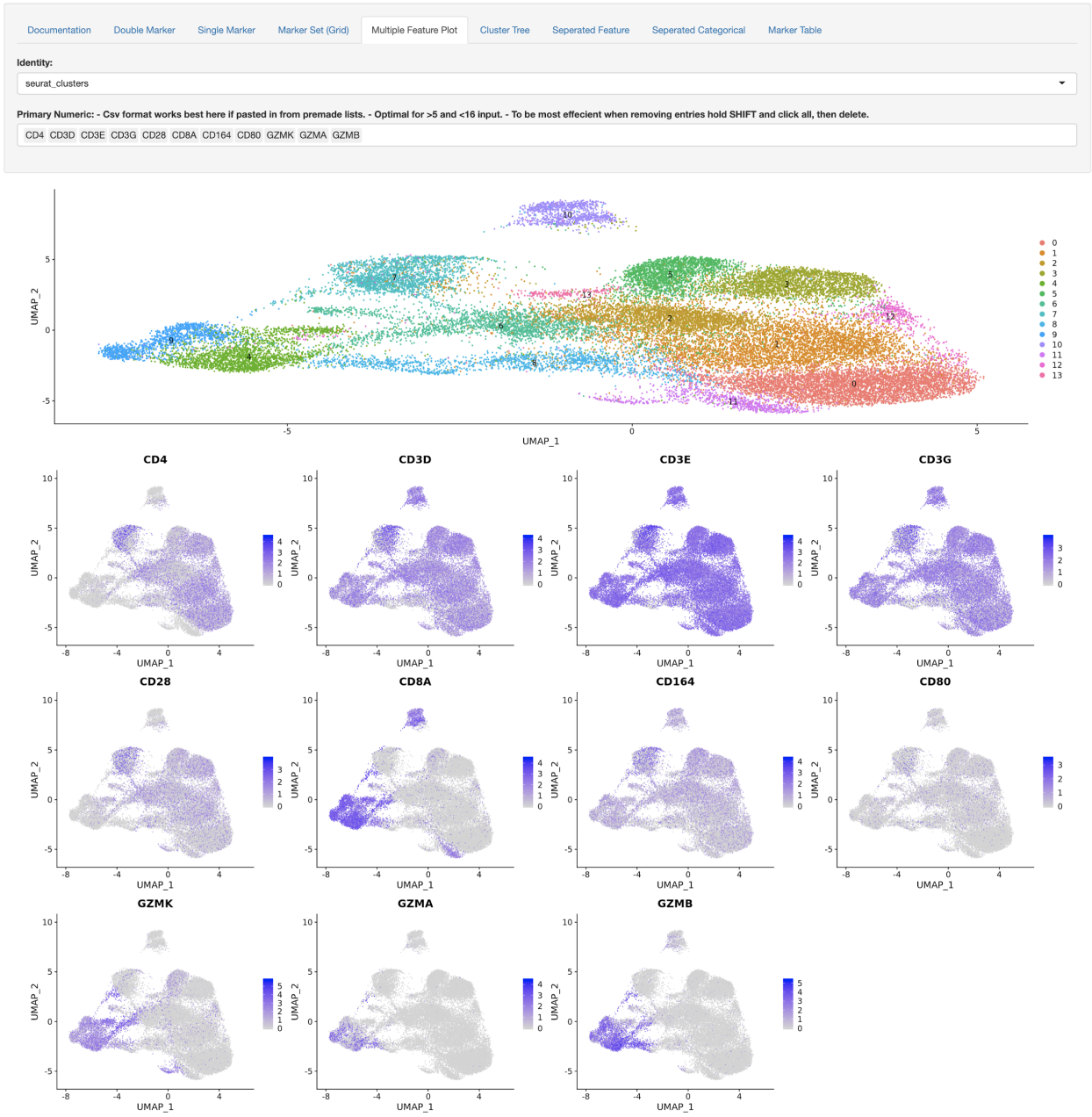


Figure 4: Multiple Feature Plot or Multiple Marker Plot.

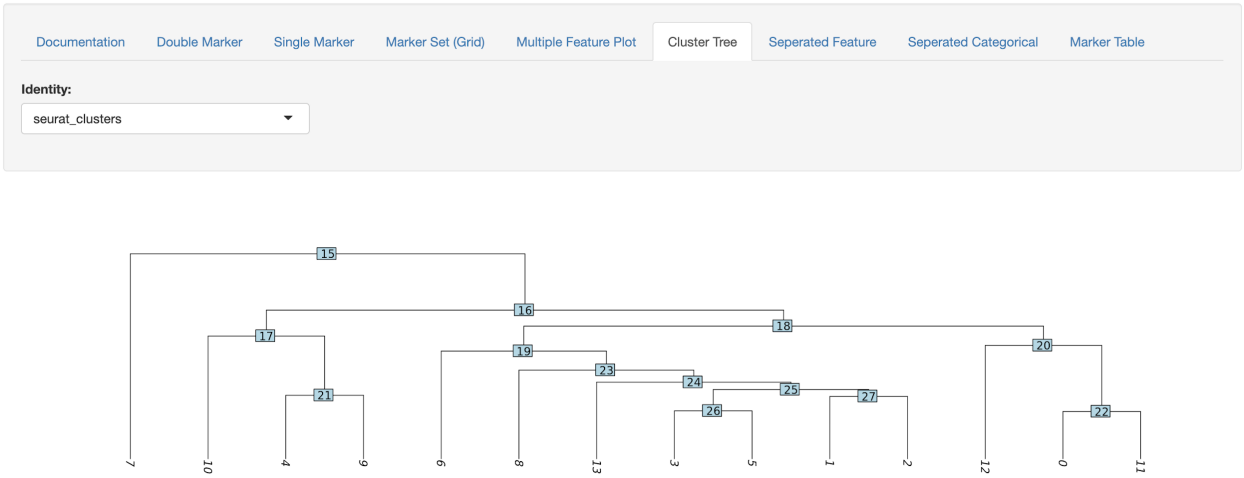
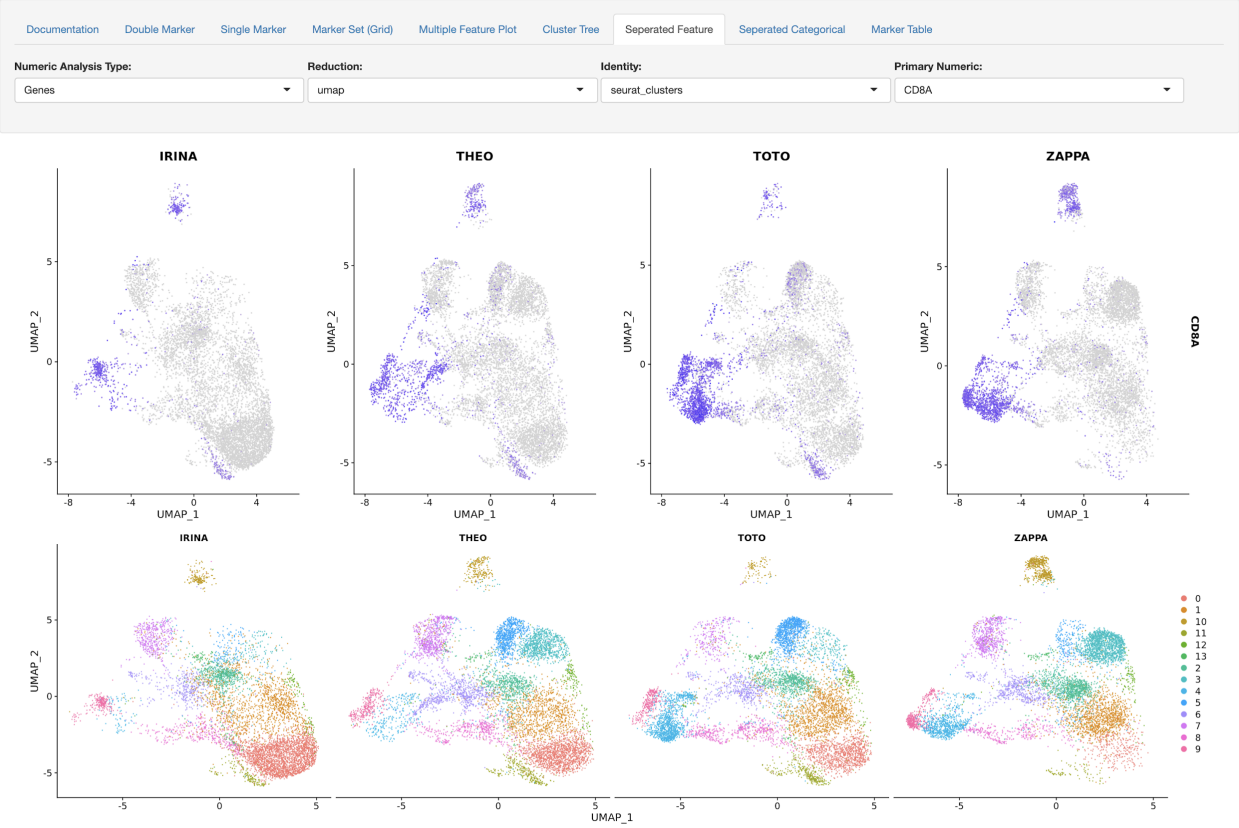


Figure 5: Cluster Tree Exploration.



	identities	IRINA.AvgExpression	THEO.AvgExpression	TOTO.AvgExpression	ZAPPA.AvgExpression	IRINA.Proportion	THEO.Proportion	TOTO.Proportion	ZAPPA.Proportion
1	0	0.0203	0.0445	0.0615	0.1127	0.3706	0.1792	0.1324	0.0394
2	1	0.0240	0.0763	0.0963	0.0729	0.2089	0.1379	0.1754	0.1656
3	10	2.2965	2.0397	2.2110	1.9640	0.0241	0.0253	0.0109	0.0676
4	11	1.1665	0.9426	1.2140	0.7989	0.0240	0.0345	0.0391	0.0105
5	12	0.2063	0.2947	0.6308	0.4298	0.0189	0.0205	0.0200	0.0148
6	13	0.0266	0.3130	0.4314	0.2094	0.0145	0.0078	0.0133	0.0075
7	2	0.0730	0.1183	0.1544	0.1031	0.0803	0.0722	0.0921	0.1232
8	3	0.0345	0.0325	0.0691	0.0293	0.0221	0.1082	0.0308	0.1810
9	4	2.5599	2.5710	2.7885	2.7840	0.0109	0.0338	0.1495	0.1017
10	5	0.1917	0.2423	0.3936	0.1970	0.0231	0.1036	0.1310	0.0372
11	6	0.2297	0.8775	0.4714	0.3064	0.0526	0.0936	0.0606	0.0672
12	7	0.1498	0.2887	0.5193	0.2477	0.0629	0.0895	0.0343	0.0834
13	8	0.1923	0.3178	0.7520	0.7007	0.0519	0.0545	0.0689	0.0432
14	9	2.7767	2.8064	2.9198	2.9077	0.0354	0.0393	0.0418	0.0578

Figure 6: Separated Feature Plots. A numeric analysis type is chosen to be either genes or metadata.

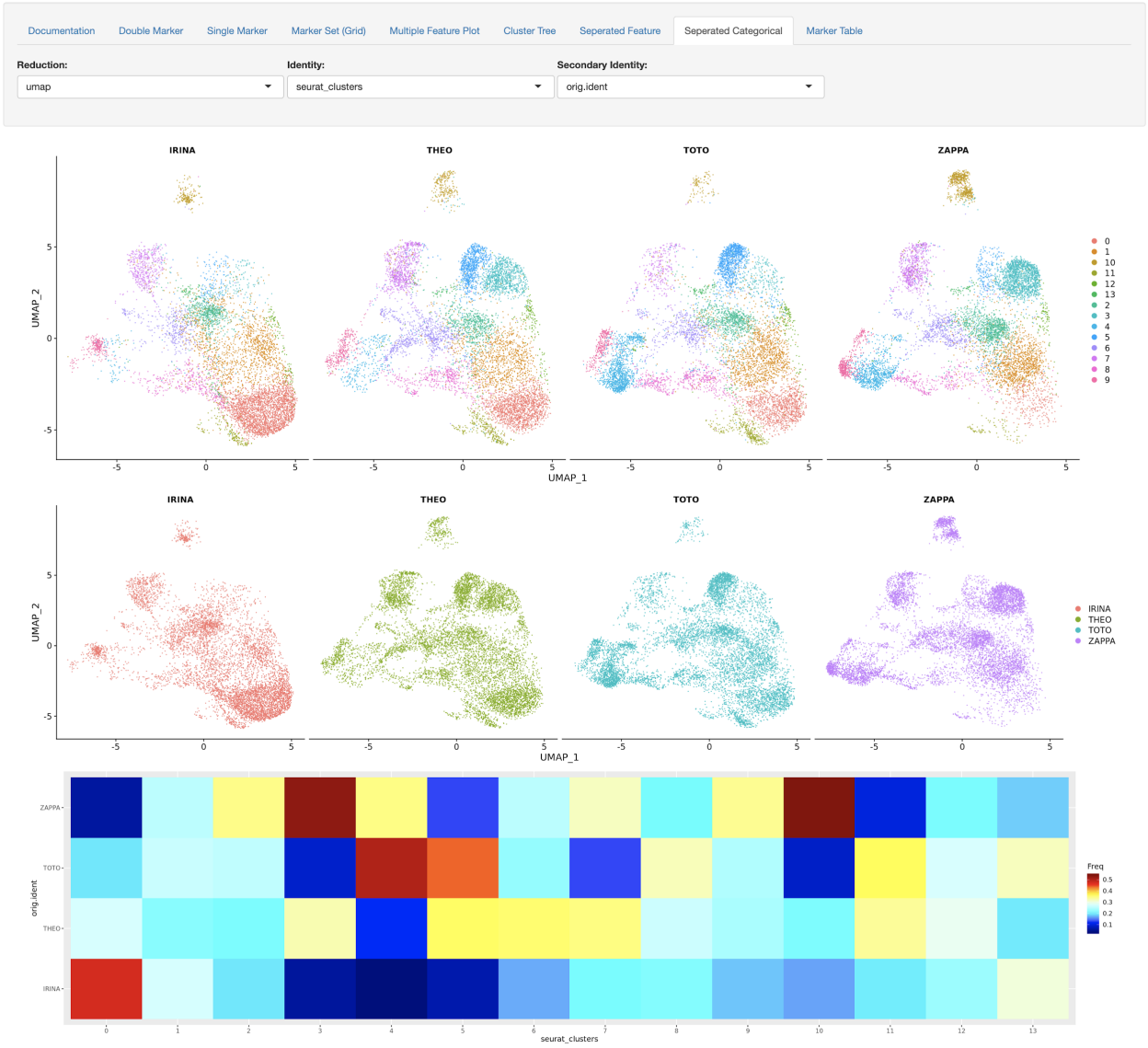


Figure 7: Separated Categorical Plot to explore the frequency of cell types in each of the samples or identities.

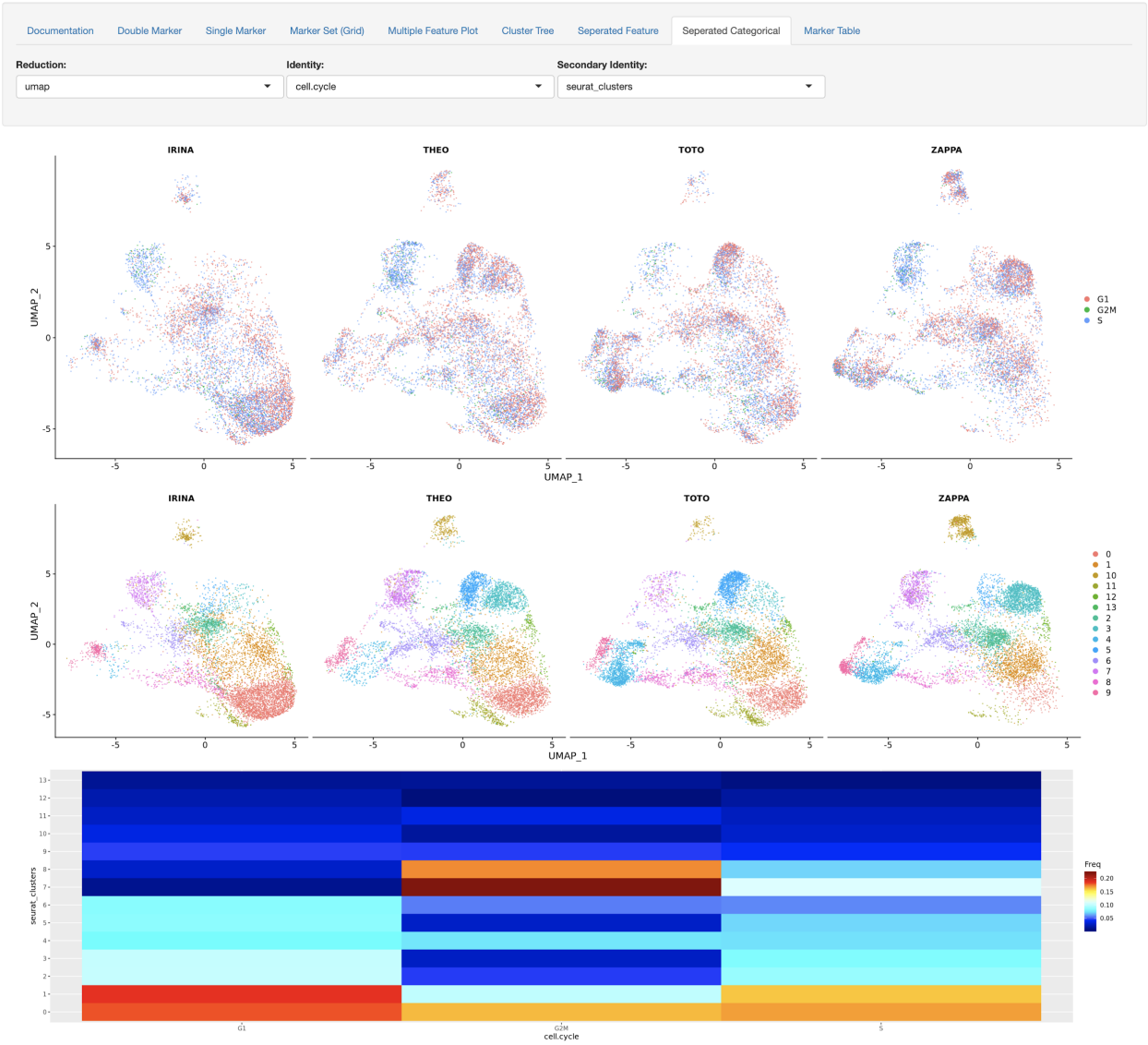


Figure 8: Separated Categorical Plots to explore the frequency of cell cycle across each of the clusters.

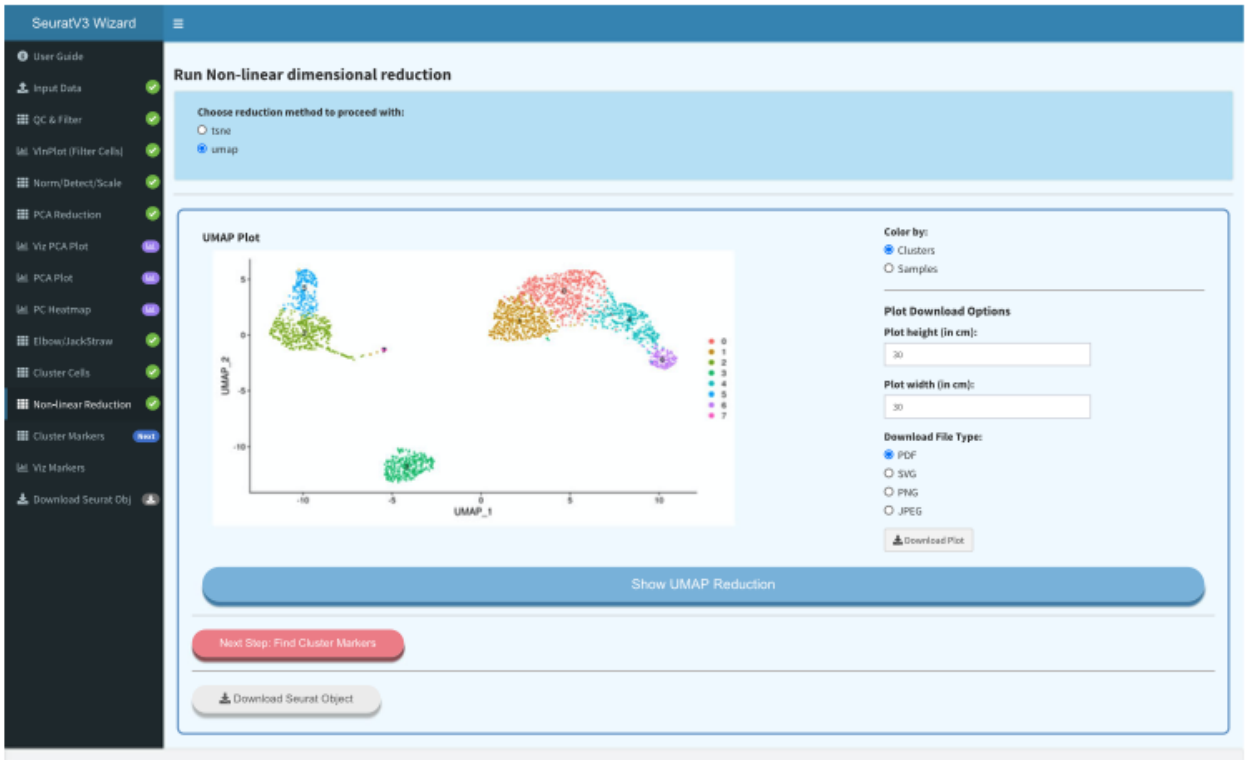


Figure 9: Visual representation of the SeuratV3 Wizard for comparison with scRNA shiny.

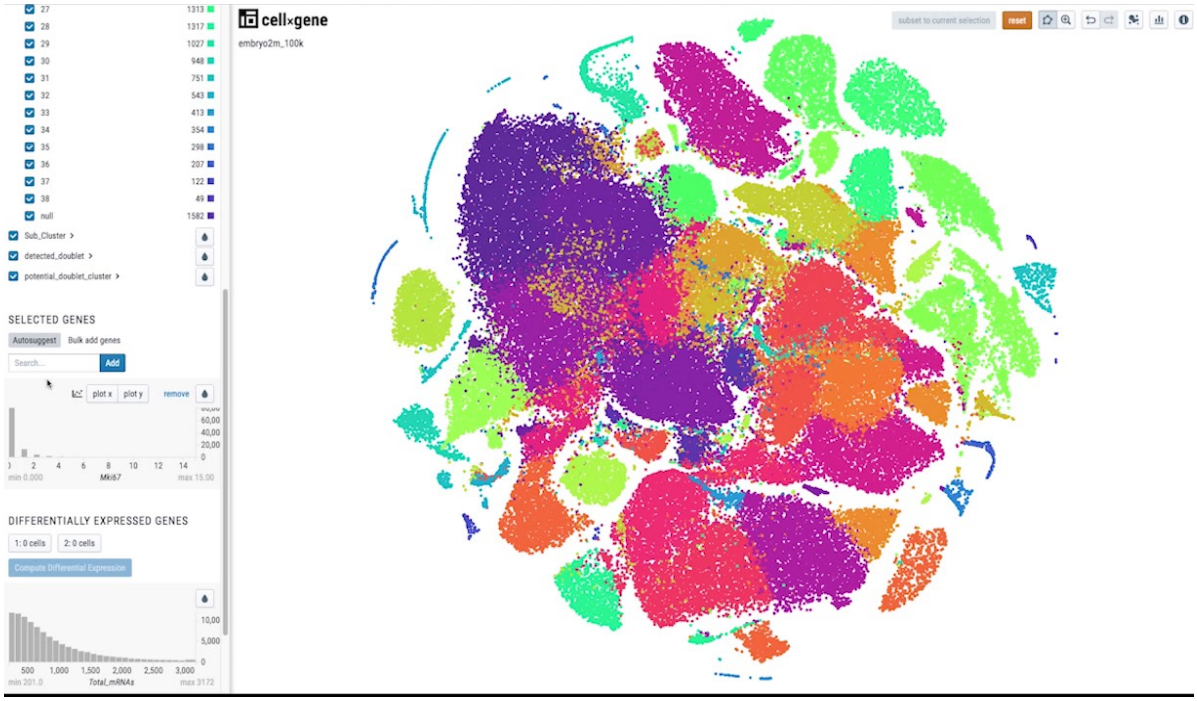


Figure 10: CellxGene representation of visualizations possibilities.

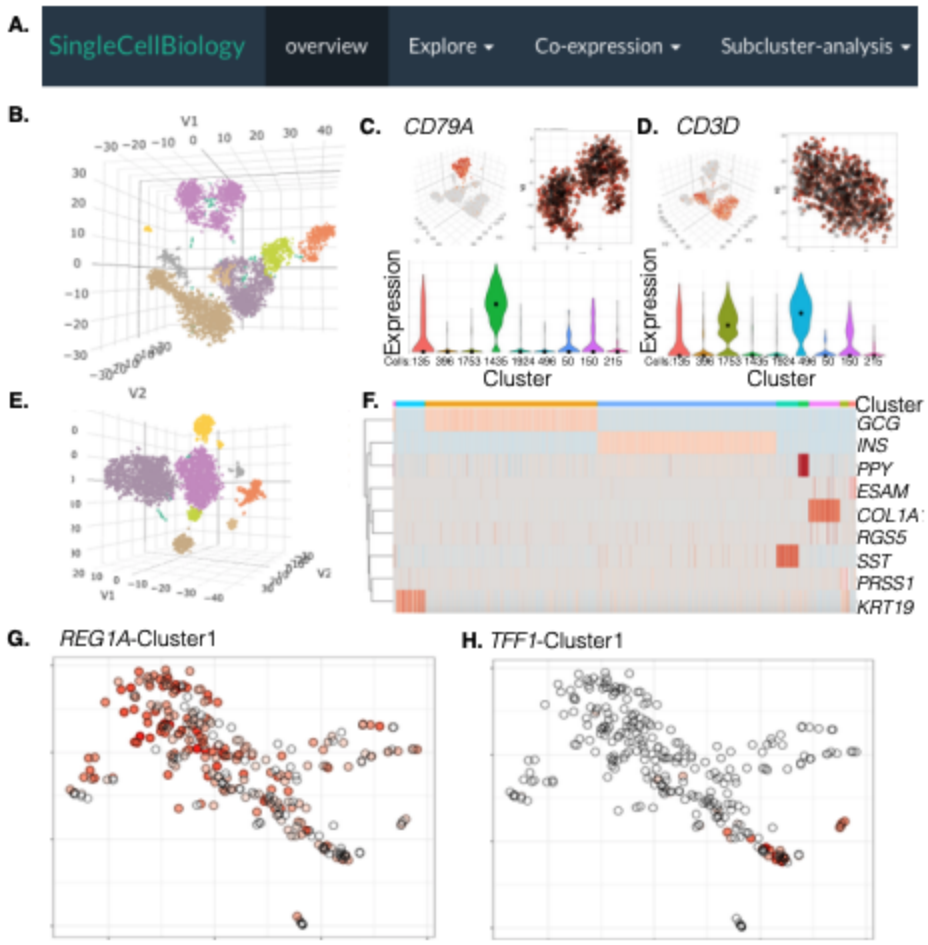


Figure 11: CellView options overview from their preprint. **A:** Demonstrates the navigation bar in the CellView application. **B,C,D,E:** Represent 3D views that are similar to the Single Marker view in scRNA Shiny. The usage of 3D UMAP and tSNE plots is not always considered a superior representation to the classical 2D plot as 3D plots are rarely acceptable in publications. **F:** Represents a similar view to what the Marker Set view in scRNA Shiny achieves. **G,H:** Finally, these figures represent a per cluster, identity analysis of expression.

DISCUSSION

Exploration of data is a key component of scRNA-seq analysis that has yet to exist in interactive forms, especially with regards to increasing experimental design complexity and analysis. Many current applications for scRNA-seq data visualization are limited. scRNA Shiny provides a simple and robust interface that is proving very helpful to researchers when initially exploring their data. scRNA Shiny can be easily deployed on personal computers using Docker systems for

across system installability and can be easily deployed on AWS for large datasets by sharing a preexisting image.

DECLARATIONS

- Ethics approval and consent to participate
    - Studies involving animals must include a statement on ethics approval.
- Consent for publication
    - Stefan Keller's Data
- Availability of data and materials
    - The software generated during and/or analysed during the current study are available in the scRNA_shiny repository, https://github.com/ucdavis-bioinformatics/scRNA_shiny
- Competing interests
    - The authors declare that they have no competing interests.
- Funding
    - Not applicable.
- Authors' contributions
    - KM wrote the software and the manuscript. MS contributed advice and intellect in the structure of the application. MB, SH contributed to testing and debugging portions of the application during usage. DH contributed to testing as well as generation of new ideas for the app. BDJ contributed to code for some portions of the application as well as consulted on display and statistics. All authors read and approved the final manuscript.
- Acknowledgements
    - Stefan Keller, for allowing his dataset to be used for explaining possible displays in the app.

CHAPTER 4: Continuation integration of diverse methodologies and some example practices.

A Summary of Data Types Discussed in CellularCall and scRNA Shiny

As discussed in the scRNA Shiny application, many quality assurance and quality control metrics performed throughout the bioinformatics processing can be evaluated using the application. This creates a seamless transition of all information from bioinformaticians to researchers and specialists. This relationship focuses on the genes and the categorical and or quantitative metadata present in the Seurat object. Metadata is typically data that is associated with each cell or barcode in single cell processes. This exists in many forms such as mitochondrial gene percentage, number of genes expressed, and total expression counts. However, there is, in theory, no limit to the amount of and types of additional metadata that can be present in the data. These typically fall into categorical and continuous metrics. For example, VDJ produces a variety of helpful information about the immune repertoire on a single-cell level and can result in a variety of metadata information that can be associated with cell types identified. One common comparison with regards to VDJ is to assess the alpha diversity of each cell type and assign those cells the diversity metrics, such as Shannon diversity, so that across sample comparisons for certain cell types can be investigated. This can similarly be performed on all clusterings performed on the data.

Expanded Data Types in Single Cell Omics Techniques

Available methods in the area of single cell genomics continue to grow. Some common ones include but are not limited to single cell ATAC-seq, single cell CHIP-seq, single cell ISO-seq, single cell CITE-seq, as well as single cell proteomic analysis. ATAC-seq is a method for assessing the accessibility of chromatin (Fang et al. 2021). CHIP-seq is a method for assessing chromatin immunoprecipitation or analyzing protein interactions with DNA (Grosselin et al. 2019). Iso-Seq reads allow for isoform analysis as they span the entire 5’ to 3’ of the transcript (Tseng and Underwood 2020). Cellular indexing of transcriptomes and epitopes by sequencing

(CITE-seq) allows for the quantification of proteins with the usage of DNA-barcoded antibodies (Stoeckius, Stoeckius, and Smibert 2017). Finally, proteomics is now being applied at the single cell level allowing for direct quantification of protein quantities at the single cell level as opposed to this being inferred by transcription (Perkel 2021).

Though these various genomic methods are different in nature, Seurat is able to handle all of these types of data as part of their data objects. These various techniques are increasingly being used in conjunction with one another, a process referred to as single cell multi-omic analysis ("Joint RNA and ATAC Analysis: 10x Multiomic" n.d.) .

Multi-Omics and the Tools CellularCall and scRNA Shiny

The number of methods available in the area of single cell genomic analysis is likely to continue to grow. This includes by is not limited to sequencing methods, cell typing methods, and databases. Visualization of these various genomic analysis techniques often take place in the same exact fashion as scRNA-seq. Typically this exists as a data matrix where rows are gene, protein, isoform counts etc. and cells being the columns which are coupled with metadata. Many of the mentioned genomic techniques can be utilized in the scRNA Shiny app in the same way it is utilized for scRNA-seq data. Due to the similarity of the data structure and the nature of generalized linear regression CellularCall can be easily applied to other forms of single cell omics data. Users can apply a markers file for the row of interest as well as identities of interest, in fact much of this information can be inferred by coupling scRNA-seq with other omic analysis of interest. Then markers for the given technology can be inferred.

Single Cell Multi-Omics: Big Data Applications

The introduction of single cell data into the area of genomics has been ground breaking for understanding the biology of cell types across tissues, diseases, and organisms. It has led to countless breakthroughs and the data generated in this space continues to grow exponentially

(Svensson, da Veiga Beltrame, and Pachter, n.d.). Aggregation of this data in the form of databases is the best long term solution for harnessing information from all existing single cell data. There exist a few in this space already, but there is much work to be done to standardize the approaches of this data

As computing resources get cheaper and standardization tools for software get better the "big data" applications will become more and more feasible. Some private industry companies have already achieved this as the scale of their data is often larger and more broadly focused in the single cell area. More public datasets will be able to achieve a similar level of comparability with the addition of more databases and standardized processing across these databases. Some currently existing databases include but are not limited to the 10X Genomics datasets, Human Cell Atlas, and JingleBells (Ner-Gaon, Melchior, and Golan 2017; Rozenblatt-Rosen et al. 2017). These databases exist as a range of options, some being standardized and other being just gateways to example datasets. It is clear that standardized databases are the future for single cell applications. As these databases become more available this will allow for large scale analysis with increased diversity of questions and hypothesis testing for researchers. In addition, this will enable large scale statistical models and machine learning methods for processes such as cell assignment. In order to get the most use of the exponentially increasing data in this area, though an ambitious task, it is important to gain this homogenization and higher level control over single cell data. This would require a valiant effort on behalf of geneticists, bioinformaticians, software engineers and other experts as well as a large amount of public funding.

REFERENCES

Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. 2019. "A Comparison of Automatic Cell Identification Methods for Single-Cell RNA Sequencing Data." *Genome Biology* 20 (1): 194.

Bolisetty, M. T., M. L. Stitzel, and P. Robson. 2017. "CellView: Interactive Exploration of High Dimensional Single Cell RNA-Seq Data." *bioRxiv*. https://www.biorxiv.org/content/10.1101/123810v1.abstract.

Cellxgene: An Interactive Explorer for Single-Cell Transcriptomics Data. n.d. Github. Accessed November 18, 2021. https://github.com/chanzuckerberg/cellxgene.

Davis, Sean. n.d. *Awesome-Single-Cell*. Github. Accessed February 5, 2021. https://github.com/seandavi/awesome-single-cell.

Fang, Rongxin, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, et al. 2021. "Comprehensive Analysis of Single Cell ATAC-Seq Data with SnapATAC." *Nature Communications* 12 (1): 1337.

Grosselin, Kevin, Adeline Durand, Justine Marsolier, Adeline Poitou, Elisabetta Marangoni, Fariba Nemati, Ahmed Dahmani, et al. 2019. "High-Throughput Single-Cell ChIP-Seq Identifies Heterogeneity of Chromatin States in Breast Cancer." *Nature Genetics* 51 (6): 1060–66.

Hafemeister, Christoph, and Rahul Satija. 2019. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." *Genome Biology* 20 (1): 296.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2020. "Integrated Analysis of Multimodal Single-Cell Data." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.10.12.335331.

"Joint RNA and ATAC Analysis: 10x Multiomic." n.d. Accessed November 19, 2021. https://satijalab.org/signac/articles/pbmc_multiomic.html.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14.

Mangul, Serghei, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. 2019. "Systematic Benchmarking of Omics Computational Tools." *Nature Communications* 10 (1): 1393.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Ner-Gaon, H., A. Melchior, and N. Golan. 2017. "Jinglebells: A Repository of Immune-Related Single-Cell Rna–sequencing Datasets." *The Journal of*. https://www.jimmunol.org/content/198/9/3375.abstract.

Perkel, Jeffrey M. 2021. "Single-Cell Proteomics Takes Centre Stage." *Nature* 597 (7877): 580–82.

Pliner, Hannah A., Jay Shendure, and Cole Trapnell. 2019. "Supervised Classification Enables Rapid Annotation of Cell Atlases." *Nature Methods* 16 (10): 983–86.

Reuell, Peter. 2018. “Harvard Teams’ Studies Featured in Science ‘Breakthrough of the Year.’” Harvard Gazette. December 21, 2018. https://news.harvard.edu/gazette/story/2018/12/studies-from-schier-lab-featured-in-science-breakthrough-of-the-year/.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. “The Human Cell Atlas: From Vision to Reality.” *Nature* 550 (7677): 451–53.

“Seurat Part 3 – Data Normalization and PCA.” 2018. January 11, 2018. https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/seurat-part-3-data-normalization/.

“Shiny.” n.d. Accessed January 31, 2021. https://shiny.rstudio.com/.

“Shiny Server.” n.d. Accessed January 31, 2021. https://rstudio.com/products/shiny/shiny-server/.

Stoeckius, Marlon, Marlon Stoeckius, and Peter Smibert. 2017. “CITE-Seq.” *Protocol Exchange*, July. https://doi.org/10.1038/protex.2017.068.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter. n.d. “A Curated Database Reveals Trends in Single-Cell Transcriptomics.” https://doi.org/10.1101/742304.

Tseng, Elizabeth, , PhD, and Jason G. Underwood , PhD. 2020. “Single-Cell Full-Length Isoform Characterization Using SMRT Sequencing.” *Genetic Engineering & Biotechnology News* 40 (3): 58–60.

Wikipedia contributors. 2020. “Multinomial Logistic Regression.” Wikipedia, The Free Encyclopedia. December 29, 2020. https://en.wikipedia.org/w/index.php?title=Multinomial_logistic_regression&oldid=996992025.

———. 2021. “T-Distributed Stochastic Neighbor Embedding.” Wikipedia, The Free Encyclopedia. January 20, 2021. https://en.wikipedia.org/w/index.php?title=T-distributed_stochastic_neighbor_embedding&oldid=1001556484.

Xu, Chen, and Zhengchang Su. 2015. “Identification of Cell Types from Single-Cell Transcriptomes Using a Novel Clustering Method.” *Bioinformatics* 31 (12): 1974–80.

Yousif, Ayman, Nizar Drou, Jillian Rowe, Mohammed Khalfan, and Kristin C. Gunsalus. 2020. “NASQAR: A Web-Based Platform for High-Throughput Sequencing Data Analysis and Visualization.” *BMC Bioinformatics* 21 (1): 267.