Salam AlAbdullatif

BF541 Internship Report

Supervisor: Dr. Joshua Campbell

Using Celda and Rmarkdown to Generate Single Cell RNA Sequencing Data Analysis Reports

During the summer, I began working at Dr. Campbell’s lab, learning the basics of single cell RNA sequencing (scRNA-seq) techniques and coding in R. The project consisted of learning how to analyze scRNA-seq data for other labs, communicating with members of those research projects about the implications of the analysis, and developing a pipeline to generate automated reports with analysis results. The purpose of this project is to collaborate with biomedical projects and to improve our ability to streamline data analysis and reduce manual input. Previously, I had no experience analyzing data in R, so my initial steps in the project were to find vignettes detailing the scRNA-seq workflow, and I quickly learned how to apply those steps to example datasets.

For a few weeks after that I became more familiar with the manual process of taking raw scRNA-seq data, applying a set of filters and quality control measures to select features which distinguish the cell clusters. Following the standard workflow for scRNA-seq analysis, I learned how to work with two different clustering packages: Seurat and Celda. The goal of both packages is identical; they aim to provide insight into the features which distinguish the cell clusters from one another and allow for classification of cell types. The downstream analysis focuses on finding those interesting features through plotting dimensionality-reduced matrices and calculating differential expression of genes.

During that time, I performed all the downstream analysis manually, selecting the number of clusters and feature modules used for clustering. I worked on a dataset of mice-derived neutrophils which revealed two large clusters distinguished by the tissue source: blood samples and bronchoalveolar lavage samples taken from inside the lungs. I met with Dr. Katrina Traber from the BU Pneumonia Biology group, discussed the background of the study, and produced a report with Rmarkdown to summarize the results of clustering and downstream analysis, identifying individual gene modules which drive the heterogeneity in gene expression between the two types of tissue in the dataset. I had back-and-forth communication directly with Dr. Traber, receiving her feedback on interesting features to improve the depth of analysis and suggestions on ways to simplify the contents of the report. As part of our discussions, I worked on creating a R shiny app which allows her and others to select individual cells on a dimensionality reduction plot and calculate differential expression on a subset of manually selected cells. This experience was particularly useful as my future goals in data analysis are to help improve and renovate a single-cell toolkit which allows users to dynamically and freely interact with the analysis reports provided to them.

The final part of the project was to integrate what I had learned in performing manual downstream analysis and Rmarkdown documents to help create an automated pipeline for the lab to use in future analysis of new datasets. The main idea behind the pipeline is to feed it the raw data and generate a full html report with all the filtering, clustering, and differential expression results. My role was to create the section of the report which performs analysis with Celda, taking an object from previous analysis sections, applying the same steps for downstream analysis as used for the neutrophil dataset, and generating all the clustering results in an organized form. We iteratively worked on this pipeline for a few weeks, collaboratively debugging and constantly improving it. Having become familiar with the datasets and the results of clustering the neutrophil dataset, it quickly became easier to understand the output from the pipeline and how to fix any problems that we came across. Also having had communication with Dr. Traber on her suggestions for the html report’s usability, I learned how to make the technical aspect of data analysis more accessible to non-bioinformatics researchers. Moving forward in my career, this project gave me very valuable experience in analyzing data, presenting it in succinct and clear fashion, receiving feedback from outside influence, and developing new software tools for collaborative programming work.

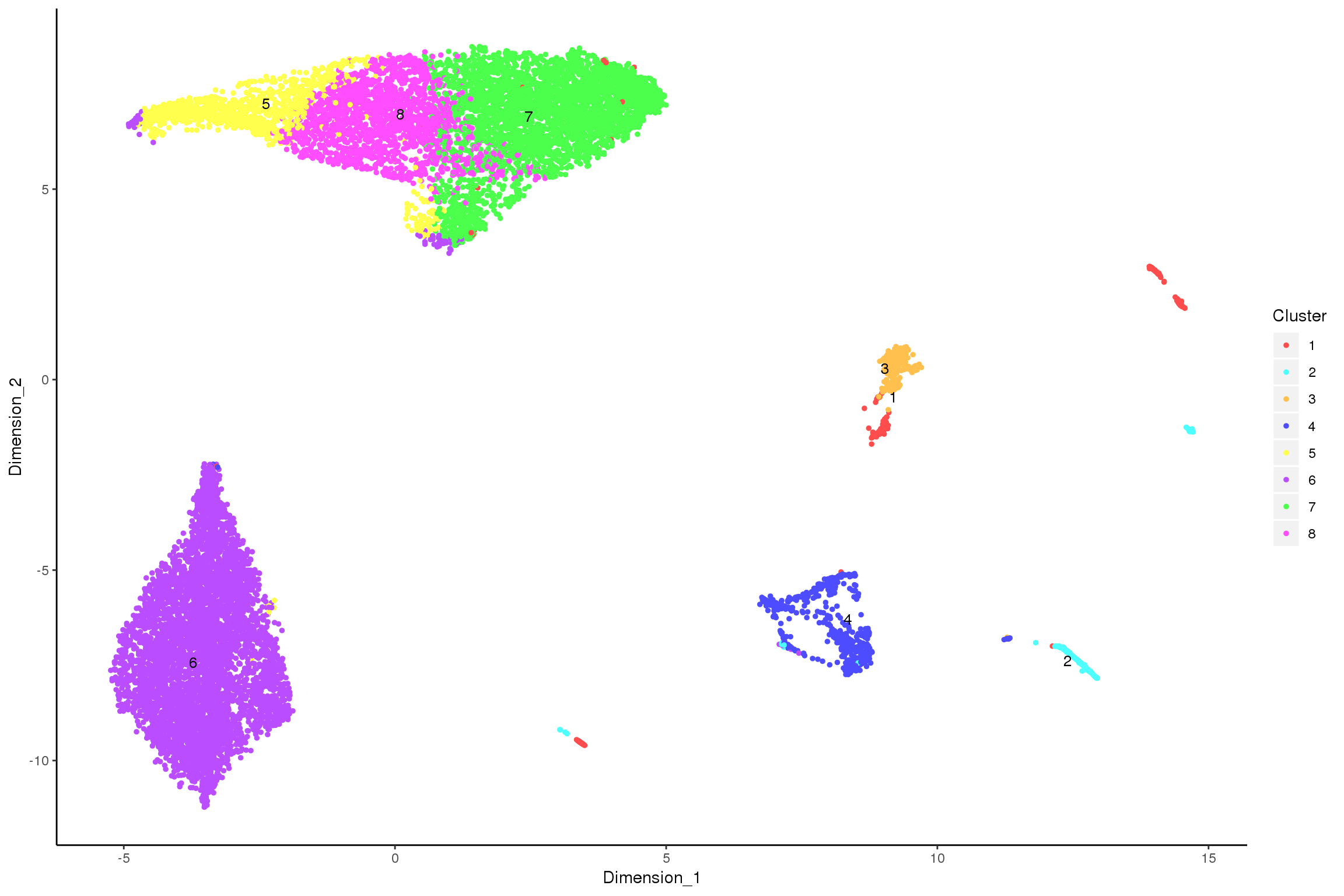


Figure 1: Cell clustering performed using Celda, represented in UMAP coordinates

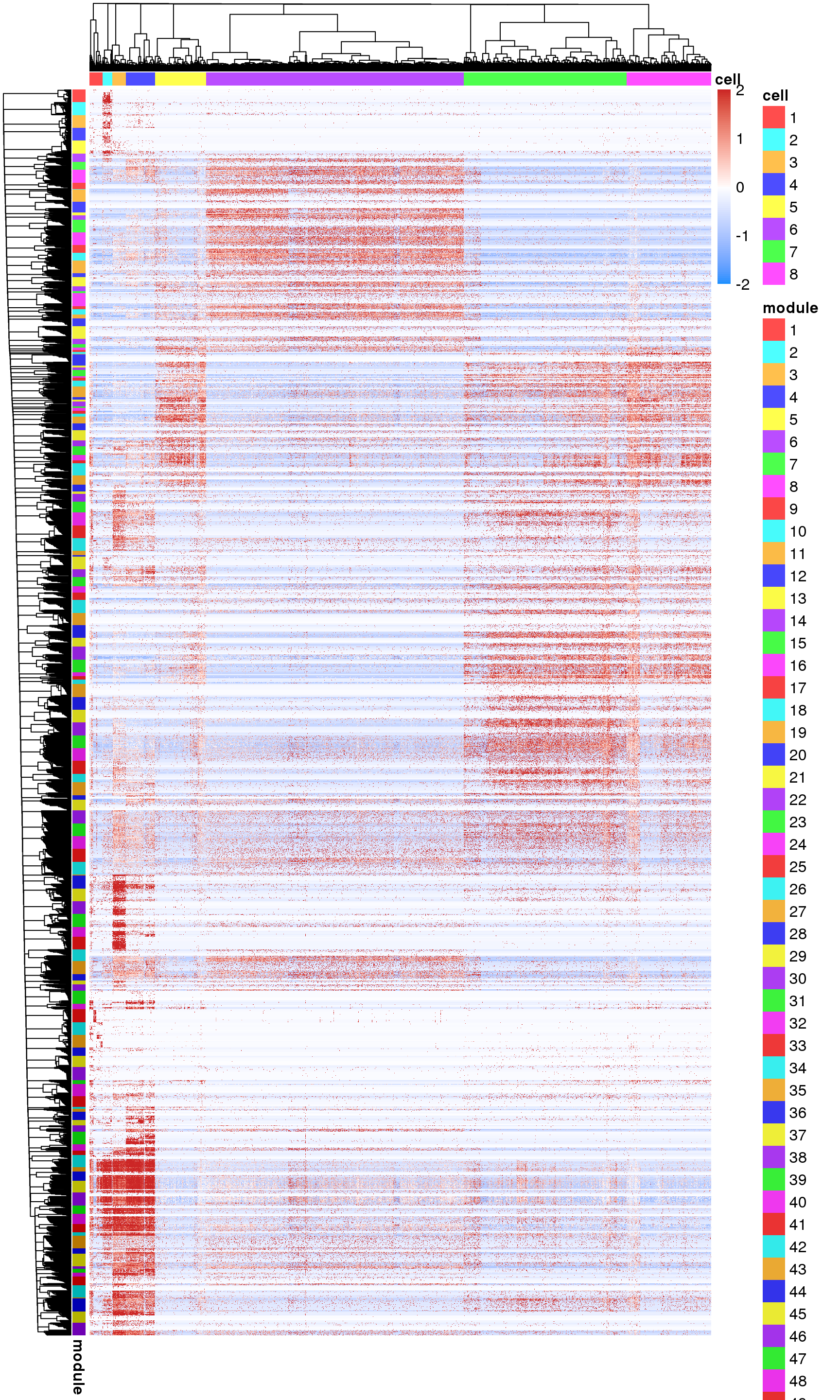


Figure 2: Heatmap of gene expression across cells, where cells are sorted by their cluster label and genes are sorted by their feature module label

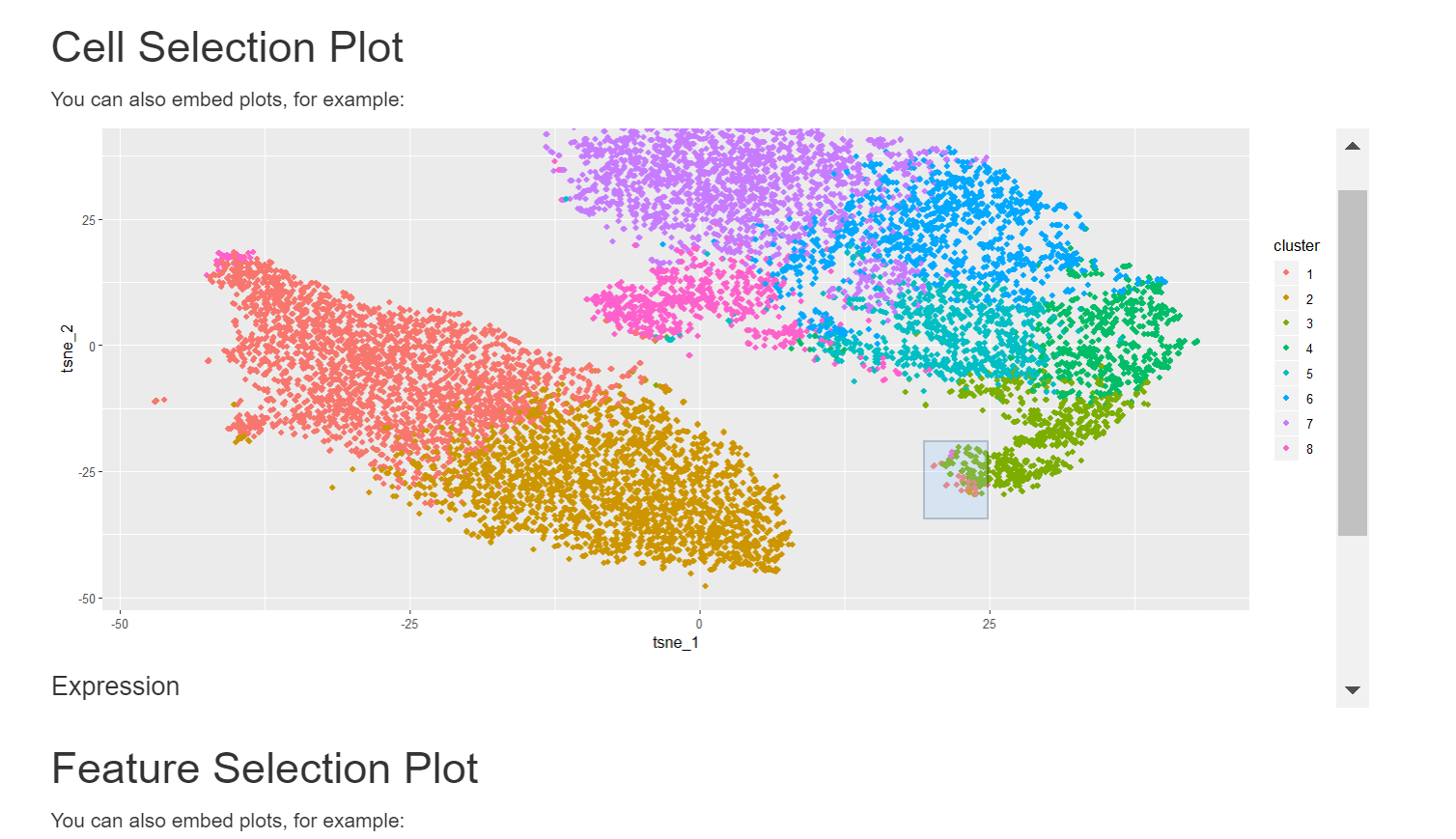


Figure 3: R shiny app created for manual selection of a subpopulation of cells, which was created for lab collaborators to freely explore the analysis of their data