# SC1: A Tool for Interactive Web-Based Single Cell RNA-Seq Data Analysis

Marmar Moussa and Ion I. Măndoiu

Computer Science & Engineering Department,
University of Connecticut, Storrs, CT 06269, USA
{marmar.moussa,ion.mandoiu}@uconn.edu

**Abstract.** Single cell RNA-seq (scRNA-Seq) is critical for studying cellular function and phenotypic heterogeneity as well as the development of tissues and tumors. Here, we present a web-based interactive scRNA-Seq data analysis tool publicly accessible at `https://sc1.engr.uconn.edu`. The tool implements a novel method of selecting informative genes based on Term-Frequency Inverse-Document-Frequency (TF-IDF) scores and provides a broad range of methods for cell clustering, differential expression, gene enrichment, interactive visualization, and cell cycle analysis. In just a few steps, researchers can generate a comprehensive initial analysis and gain powerful insights from their single cell RNA-seq data.

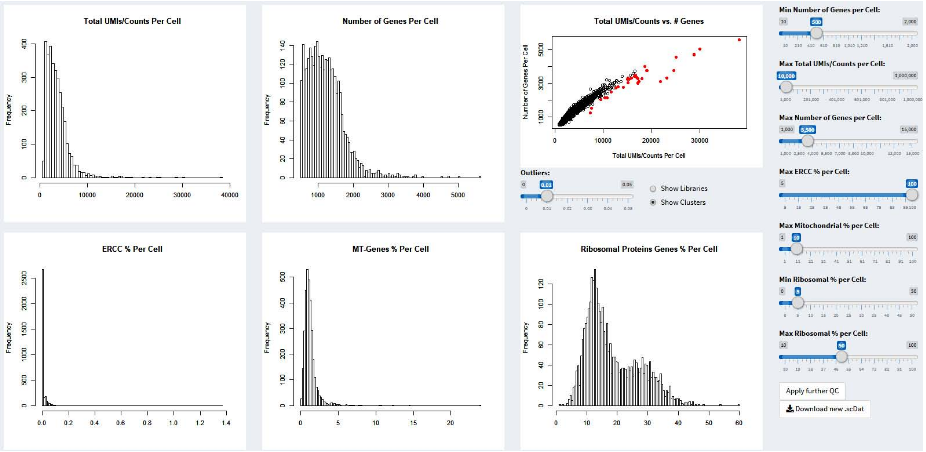
## 1 Introduction

Currently there are only few packages for comprehensive scRNA-Seq data analysis. Most of them are implemented using the R programming language, require considerable programming knowledge, and are not easy to use by researchers in life sciences.

In this work, we present a web-based, highly interactive scRNA-Seq data analysis tool publicly accessible at `https://sc1.engr.uconn.edu`. The tool includes several data quality control (QC) options, a novel method for gene selection based on *Term-Frequency Inverse-Document-Frequency (TF-IDF)* scores [9], followed by cell clustering and visualization tools as well as Differential Expression (DE) analysis and gene enrichment steps. Additional analyses include various 3D interactive visualizations based on t-SNE and UMAP dimensionality reduction algorithms as well as a novel approach to clustering and ordering cells according to their cell cycle phase [7]. With robust default parameter values SC1 empowers researchers to generate a comprehensive initial analysis of their scRNA-Seq data in just a few steps, while also allowing them to conduct in depth interactive data exploration and parameter tuning.

## 2 SC1 Workflow

The SC1 workflow is implemented in the R programming language, with an interactive web-based front-end built using the Shiny framework [1]. In the following we present details of the main analysis steps of the workflow.



Outliers: 0.01 (0 – 0.05)

○ Show Libraries
● Show Clusters

Min Number of Genes per Cell: 500 (10 – 2,000)

Max Total UMIs/Counts per Cell: 10,000 (1,000 – 1,000,000)

Max Number of Genes per Cell: 3,500 (1,000 – 15,000)

Max ERCC % per Cell: 100 (5 – 100)

Max Mitochondrial % per Cell: 10 (1 – 100)

Min Ribosomal % per Cell: 5 (0 – 50)

Max Ribosomal % per Cell: 50 (10 – 100)

**Fig. 1.** SC1 QC dashboard.

*Data Pre-Processing.* Before a detailed analysis of scRNA-Seq datasets (in 10X Genomics or csv format) can be performed, several pre-processing steps are carried out, starting with an initial quality control step in which cells with less than 500 detected genes and genes detected in less that 10 cells are excluded. Imputation is provided as an optional pre-processing step. Empirical experiments in [6] show that over-imputation is a concern for most existing methods. In SC1 we implemented the Locality Sensitive Imputation method (LSImpute) from [8], which was shown in [6] to yield high accuracy with minimum over-imputation. SC1 pre-processing also includes performing dimensionality reduction using three commonly used algorithms: Principal Component Analysis (PCA) [2], t-distributed Stochastic Neighborhood Embedding (t-SNE) projections [11], and Uniform Manifold Approximation and Projection (UMAP) [5].

*scDat Upload.* Pre-processed data is saved in SC1's ".scDat" file format that can then be uploaded for interactive analysis. Several publicly available datasets from [4], [12] and [3] spanning different scRNA-Seq technologies are provided in SC1 as example datasets. Initial data exploration includes detecting the species (mouse or human), generating basic summary statistics including the number of expressed genes and the number of cells per library, and the ability to relabel the libraries. 'At-a-glance' two dimensional views of the data are also generated based on PCA, tSNE, and UMAP.

*Quality Control Dashboard.* Before further analyses, SC1 allows users to perform additional Quality Control (QC) checks as shown in Figure 1, whereby poor quality cells and outlier cells and genes can be excluded from subsequent analysis. The tool implements widely used criteria for cell filtering: library size, number of detected genes, as well as the fraction of reads mapping to mitochondrial genes, ribosomal protein genes, or synthetic spike-ins. SC1 also allows outlier removal

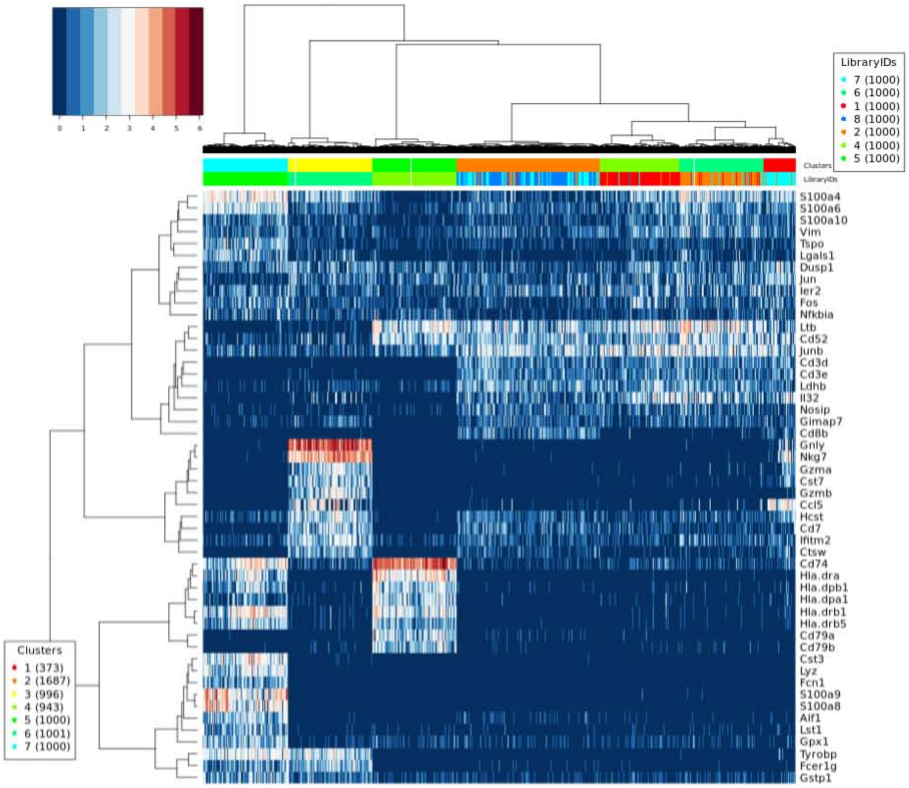
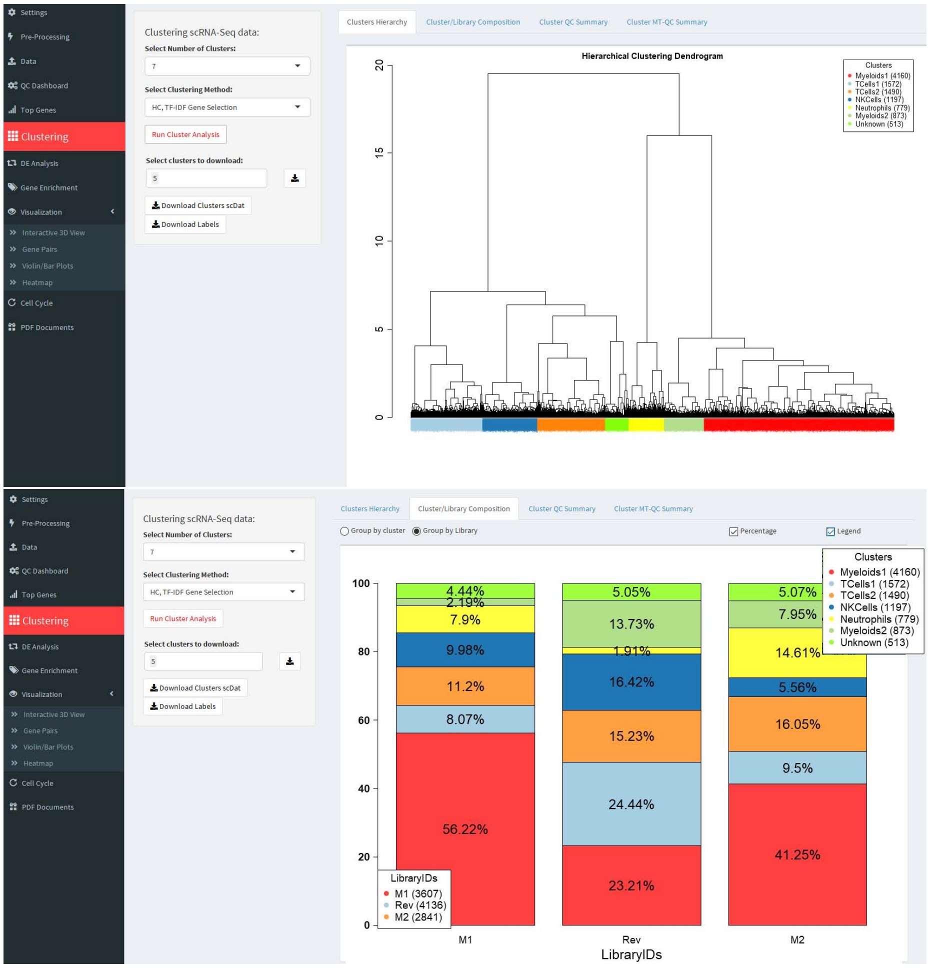


**Fig. 2.** Heat map of genes with top average TF-IDF scores for cells of the 7-class PBMC mixture from [9].

based on the ratio between the number of detected genes to total read/UMI count per cell.

*Gene Selection.* SC1 implements a novel method of selecting informative genes based on the average TF-IDF (*Term Frequency times Inverse Document Frequency*) scores, as detailed in [9]. TF-IDF scores are applied to scRNA-Seq data by considering the cells to be analogous to documents; in this analogy, genes correspond to words and UMI counts replace word counts. The TF-IDF scores can then be computed from UMI counts (or expression values). Similar to document analysis, the genes with highest TF-IDF scores in a cell are expected to provide most information about the cell type. Genes with highest average TF-IDF scores differentiate best between heterogeneous cell populations; visually this leads to a clear "chess-board" effect in the heat map constructed using the top average TF-IDF genes as shown in Fig. 2.

*Clustering.* By default, SC1 automatically infers the number of clusters using the Gap Statistics method as described in [9]. However, users can also manually



Clustering scRNA-Seq data:

Select Number of Clusters: 7

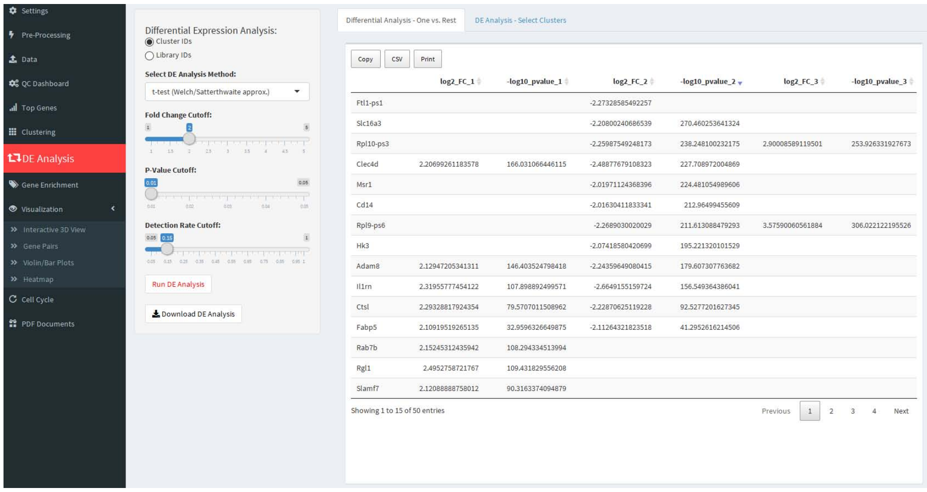
Select Clustering Method: HC, TF-IDF Gene Selection

Select clusters to download: 5

**Fig. 3.** SC1 clustering.

specify the number of clusters based on prior knowledge of the expected sample heterogeneity. Valuable insight into sample heterogeneity is also provided by inspecting the heat map generated using the top TF-IDF genes (Fig. 2) before clustering. Clustering can be performed using Ward's Hierarchical Agglomerative Clustering or Spherical K-means (both using the top average TF-IDF genes as features) or using Graph-based Clustering using binarized TF-IDF data as described in [9]. Several visualizations describe clustering details (see Fig. 3).

*Differential Expression Analysis.* Differential expression (DE) analysis is done by performing "One vs. the Rest" t-tests for each of the identified clusters. Results of the Log2 Fold Change and the p-value from the analysis are provided as a downloadable numeric matrix. A custom test of two selected groups of clusters or



Differential Expression Analysis:
● Cluster IDs
○ Library IDs

Select DE Analysis Method: t-test (Welch/Satterthwaite approx.)

Fold Change Cutoff: 2

P-Value Cutoff: 0.01

Detection Rate Cutoff: 0.15

| | log2_FC_1 | -log10_pvalue_1 | log2_FC_2 | -log10_pvalue_2 | log2_FC_3 | -log10_pvalue_3 |
|---|---|---|---|---|---|---|
| Ftl1-ps1 | | | -2.27328585492257 | | | |
| Slc16a3 | | | -2.20800240686539 | 270.460253641324 | | |
| Rpl10-ps3 | | | -2.25987549248173 | 238.248100232175 | 2.90008589119501 | 253.926331927673 |
| Clec4d | 2.20699261183578 | 166.031066446115 | -2.48877679108323 | 227.708972004869 | | |
| Msr1 | | | -2.01971124368396 | 224.481054989606 | | |
| Cd14 | | | -2.01630411833341 | 212.96499455609 | | |
| Rpl9-ps6 | | | -2.2689030020029 | 211.613088479293 | 3.57590060561884 | 306.022122195526 |
| Hk3 | | | -2.07418580420699 | 195.221320101529 | | |
| Adam8 | 2.12947205341311 | 146.403524798418 | -2.24359649080415 | 179.607307763682 | | |
| Il1rn | 2.31955777454122 | 107.898892499571 | -2.6649155159724 | 156.549364386041 | | |
| Ctsl | 2.29328817924354 | 79.5707011508962 | -2.22870625119228 | 92.5277201627345 | | |
| Fabp5 | 2.10919519265135 | 32.9596326649875 | -2.11264321823518 | 41.2952616214506 | | |
| Rab7b | 2.15245312435942 | 108.294334513994 | | | | |
| Rgl1 | 2.4952758721767 | 109.431829556208 | | | | |
| Slamf7 | 2.12088888758012 | 90.3163374094879 | | | | |

Showing 1 to 15 of 50 entries

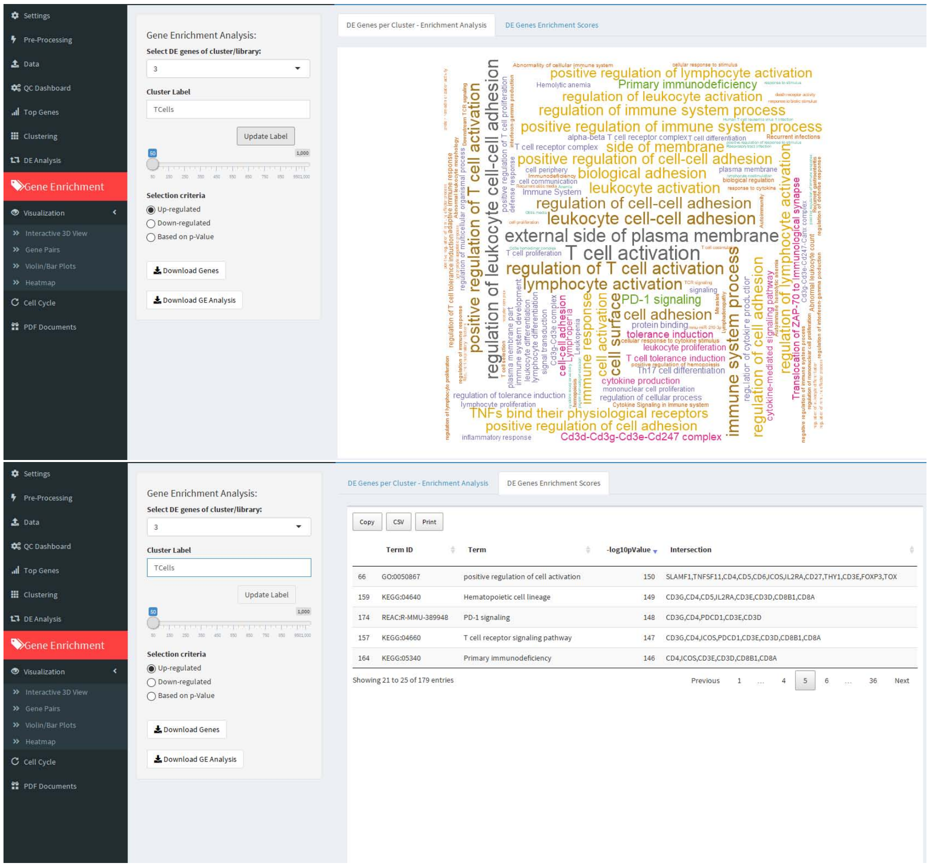
**Fig. 4.** SC1 differential expression analysis.

libraries is also provided, with results provided both as a downloadable numeric table and as a Volcano plot visualizing the Log2 Fold Change and p-values for the tested groups (Fig. 4).

*Enrichment Analysis* DE analysis is followed by cluster-based gene functional enrichment analysis performed using the 'gProfileR' R package [10] with results visualized as word clouds (Fig. 5) and provided as downloadable term significance values to help with cluster annotation. Labels assigned to the clusters at this step update throughout SC1 tool output and visualization plots.

*Interactive Data Visualization.* Many SC1 analysis steps generate visualizations of the results, including for instance the violin plots showing the probability density of gene expression values for each selected cluster/library and the bar-plots showing percentage of cells expressing selected genes by cluster or by library. Additional visualizations include:

- Clustering and gene co-expression visualization. SC1 includes multiple interactive visualization options; the interactive 3D t-SNE or UMAP visualization tabs include the ability to select genes individually, in pairs, or in groups as predefined gene sets. Cells are identified where all (AND) or any (OR) of the selected genes are detected. Identified cell populations can be selected or excluded to form a subset that can be downloaded and used to form a new sub-population for further analysis in SC1 (Fig. 6). Identifying various cell populations in SC1 and downloading relevant cells' expression profiles can be achieved in various ways in SC1: by selecting pre-defined libraries or conditions or selecting cell populations based on gene selection, also selecting specific cell types from clustering analysis results. Gene pair co-expression


Settings
Pre-Processing
Data
QC Dashboard
Top Genes
Clustering
DE Analysis
Gene Enrichment
Visualization
Interactive 3D View
Gene Pairs
Violin/Bar Plots
Heatmap
Cell Cycle
PDF Documents
DE Genes per Cluster - Enrichment Analysis
DE Genes Enrichment Scores
Update Label
Download Genes
Download GE Analysis
Copy
CSV
Print
positive regulation of lymphocyte activation
Primary immunodeficiency
regulation of leukocyte activation
regulation of immune system process
positive regulation of immune system process
side of membrane
positive regulation of cell-cell adhesion
biological adhesion
leukocyte activation
regulation of cell-cell adhesion
leukocyte cell-cell adhesion
external side of plasma membrane
T cell activation
regulation of T cell activation
lymphocyte activation
PD-1 signaling
cell adhesion
TNFs bind their physiological receptors
positive regulation of cell adhesion
Cd3d-Cd3g-Cd3e-Cd247 complex
regulation of leukocyte cell-cell adhesion
positive regulation of T cell activation
immune system process
regulation of cell adhesion
regulation of lymphocyte activation
immune response
cell activation
cell surface

Gene Enrichment Analysis:

Select DE genes of cluster/library: 3

Cluster Label: TCells

Selection criteria: ● Up-regulated ○ Down-regulated ○ Based on p-Value

| | Term ID | Term | -log10pValue | Intersection |
|---|---|---|---|---|
| 66 | GO:0050867 | positive regulation of cell activation | 150 | SLAMF1,TNFSF11,CD4,CD5,CD6,ICOS,IL2RA,CD27,THY1,CD3E,FOXP3,TOX |
| 159 | KEGG:04640 | Hematopoietic cell lineage | 149 | CD3G,CD4,CD5,IL2RA,CD3E,CD3D,CD8B1,CD8A |
| 174 | REAC:R-MMU-389948 | PD-1 signaling | 148 | CD3G,CD4,PDCD1,CD3E,CD3D |
| 157 | KEGG:04660 | T cell receptor signaling pathway | 147 | CD3G,CD4,ICOS,PDCD1,CD3E,CD3D,CD8B1,CD8A |
| 164 | KEGG:05340 | Primary immunodeficiency | 146 | CD4,ICOS,CD3E,CD3D,CD8B1,CD8A |

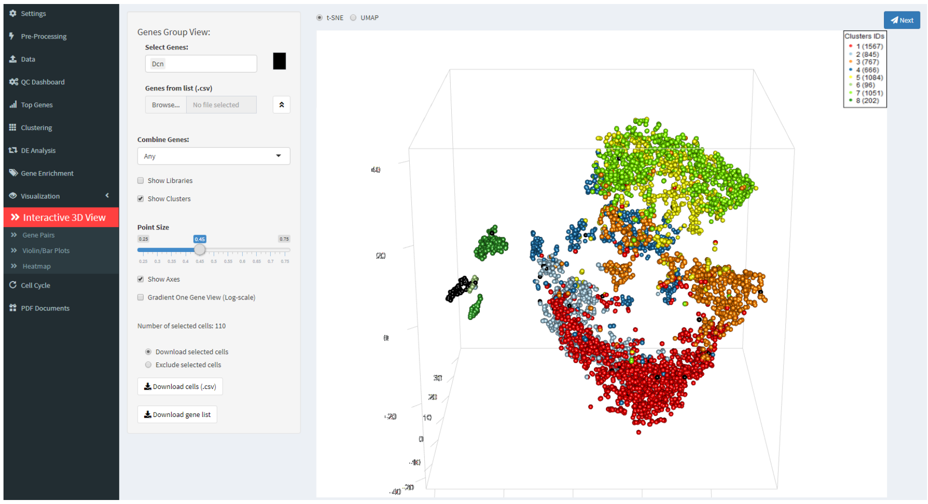
Showing 21 to 25 of 179 entries

Previous 1 … 4 5 6 … 36 Next

**Fig. 5.** SC1 cluster-based gene enrichment analysis.

can also be visualized using interactive 3D plots as well as scatter plots Fig. 7)

– Detailed and summary heatmaps. SC1 provides several ways to select genes and cells visualized in configurable heat maps. Automatic identification of informative genes based on average TF-IDF allows the generation of exploratory heat maps to investigate the heterogeneity of the data. Also, a list of highly expressed/abundant genes can be downloaded from SC1 and used to construct a heat map. SC1 also supports custom gene selection by manually selecting or uploading a list of genes of interest to use for heat map construction. After the DE analysis step is concluded, the list of differentially expressed genes can also be visualized as a heat map. The expression/count values are by default log transformed in SC1 heat maps using the $log2(x + 1)$ transformation. The summary heat map view in SC1 provides a “pseudo-bulk” view of the data, showing average expression profiles for selected genes by cluster or library (Fig. 8). The gene expression levels



Genes Group View:

Select Genes: Dcn

Genes from list (.csv): No file selected

Combine Genes: Any

☐ Show Libraries

☑ Show Clusters

Point Size: 0.45 (0.25 – 0.75)

☑ Show Axes

☐ Gradient One Gene View (Log-scale)

Number of selected cells: 110

● Download selected cells

○ Exclude selected cells

● t-SNE ○ UMAP

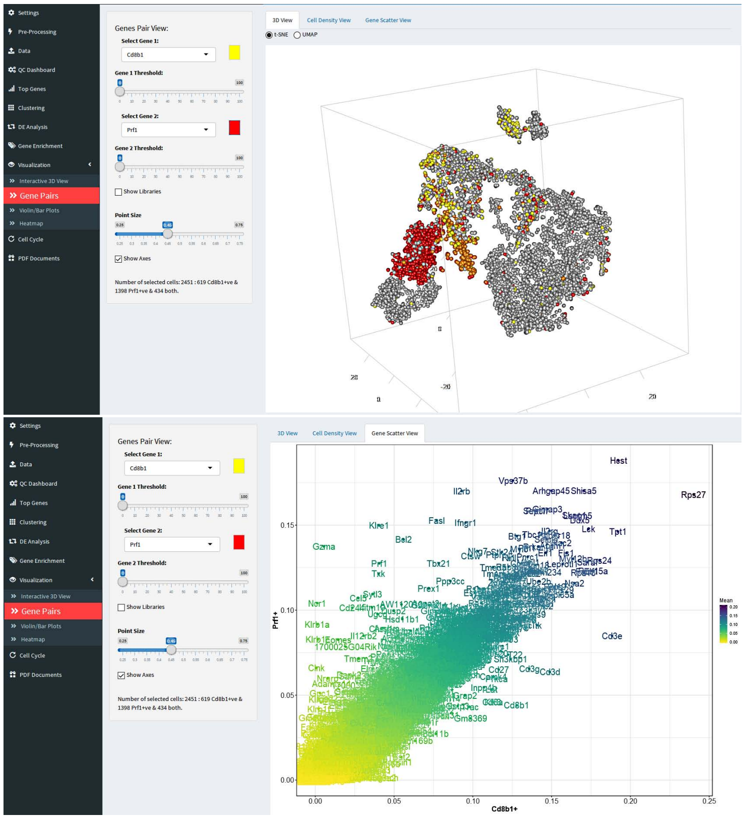
**Fig. 6.** SC1 3D visualization of clustering results and selected genes on data from [4].

in summary heat maps are row-normalized, i.e., gene means expressions in libraries and clusters are normalized by dividing by the max mean expression of each gene over all libraries and clusters. This assigns a maximum value of 1 (red) to the groups for which the mean expression of the gene is the highest.

*Cell Cycle Analysis.* The variation in the gene expression profiles of single cells in different phases of the cell cycle can present a leading source of variance between cells and can interfere with cell type identification and functional analysis of scRNA-Seq data. In SC1, an orthogonal analysis of cell cycle effects can be performed at any stage of the analysis by clustering and ordering cells according to the expression levels of cell cycle genes, as described in [7].

## 3 Conclusion

SC1 provides a powerful tool for interactive web-based analysis of scRNA-Seq data. The SC1 workflow is implemented in the R programming language, with an interactive web-based front-end built using the Shiny framework [1]. SC1 employs a novel method for gene selection based on *Term-Frequency Inverse-Document-Frequency (TF-IDF)* scores [9], and provides a broad range of methods for cell clustering, differential expression analysis, gene enrichment, visualization, and cell cycle analysis. Future work includes integrating additional clustering methods, as well as other differential expression analysis methods and integrating methods for cell differentiation analysis. As the amount of scRNA-Seq data continues to grow at an accelerated pace, we hope that SC1 will help researchers to fully leverage the power of this technology to gain novel biological insights.



Genes Pair View:

Select Gene 1: Cd8b1 🟨

Gene 1 Threshold: 0

Select Gene 2: Prf1 🟥

Gene 2 Threshold: 0

☐ Show Libraries

Point Size 0.45

☑ Show Axes

Number of selected cells: 2451 : 619 Cd8b1+ve & 1398 Prf1+ve & 434 both.

Genes Pair View:

Select Gene 1: Cd8b1 🟨

Gene 1 Threshold: 0

Select Gene 2: Prf1 🟥

Gene 2 Threshold: 0

☐ Show Libraries

Point Size 0.45

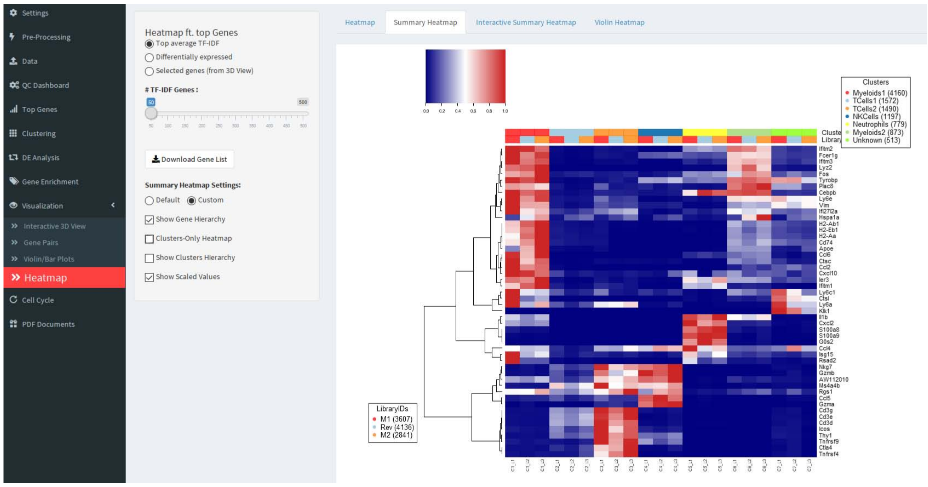
☑ Show Axes

Number of selected cells: 2451 : 619 Cd8b1+ve & 1398 Prf1+ve & 434 both.

**Fig. 7.** SC1 gene co-expression visualization.

## References

1. W Chang, J Cheng, JJ Allaire, Y Xie, and J McPherson. Shiny: Web application framework for R. *http://CRAN.R-project.org/package=shiny*, 2017.
2. N Benjamin Erichson, Sergey Voronin, Steven L Brunton, and J Nathan Kutz. Randomized matrix decompositions using R. *arXiv preprint arXiv:1608.02148*, 2016.
3. Matthew M Gubin, Ekaterina Esaulova, Jeffrey P Ward, Olga N Malkova, Daniele Runci, Pamela Wong, Takuro Noguchi, Cora D Arthur, Wei Meng, Elise Alspach, et al. High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. *Cell*, 175(4):1014–1030, 2018.
4. Samuel W Lukowski, Zewen K Tuong, Katharina Noske, Anne Senabouth, Quan H Nguyen, Stacey B Andersen, H Peter Soyer, Ian H Frazer, and Joseph E Powell. Detection of HPV E7 transcription at single-cell resolution in epidermis. *Journal of Investigative Dermatology*, 138(12):2558–2567, 2018.



Heatmap ft. top Genes

● Top average TF-IDF

○ Differentially expressed

○ Selected genes (from 3D View)

**# TF-IDF Genes :** 50 (50–500)

Download Gene List

**Summary Heatmap Settings:**

○ Default ● Custom

☑ Show Gene Hierarchy

□ Clusters-Only Heatmap

□ Show Clusters Hierarchy

☑ Show Scaled Values

**Fig. 8.** Summary heat map showing cluster/library breakdown mean expression profiles of selected genes.

5. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
6. M. Moussa and I.I. Măndoiu. Locality sensitive imputation for single-cell RNA-Seq data. *Journal of Computational Biology*, 26, 2019.
7. Marmar Moussa. Computational cell cycle analysis of single cell RNA-Seq data. In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 1–1. IEEE, 2018.
8. Marmar Moussa and Ion Măndoiu. Locality sensitive imputation for single-cell RNA-Seq data. *ISBRA2018 Proceedings*, 2018.
9. Marmar Moussa and Ion Măndoiu. Single cell RNA-Seq data clustering using TF-IDF based methods. *BMC Genomics*, 19(Suppl 6):4922, 2018.
10. Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2):W193–W200, 2007.
11. L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
12. Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *bioRxiv*, page 065912, 2016.