2021-08-13

Sciviewer enables interactive visual interrogation of single-cell RNA-Seq data from the Python programming environment [preprint]

Dylan Kotliar
Massachusetts Institute of Technology

Et al.

Title

Sciviewer enables interactive visual interrogation of single-cell RNA-Seq data from the Python programming environment

Authors

Dylan Kotliar[1,2], Andres Colubri[2,3]

Affiliations

[1]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. 3 Programming in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655

Abstract

Summary: Visualizing two-dimensional (2D) embeddings (e.g. UMAP or tSNE) is a key step in interrogating single-cell RNA sequencing (scRNA-Seq) data. Subsequently, users typically iterate between programmatic analyses (e.g. clustering and differential expression) and visual exploration (e.g. coloring cells by interesting features) to uncover biological signals in the data. Interactive tools exist to facilitate visual exploration of embeddings such as performing differential expression on user-selected cells. However, the practical utility of these tools is limited because they don't support rapid movement of data and results to and from the programming environments where the bulk of data analysis takes place, interrupting the iterative process. Here, we present the Single-cell Interactive Viewer (Sciviewer), a tool that overcomes this limitation by allowing interactive visual interrogation of embeddings from within Python. Beyond differential expression analysis of user-selected cells, Sciviewer implements a novel method to identify genes varying locally along any user-specified direction on the embedding. Sciviewer enables rapid and flexible iteration between interactive and programmatic modes of scRNA-Seq exploration, illustrating a useful approach for analyzing high-dimensional data.

Availability and implementation: Code and examples are provided at https://github.com/colabobio/sciviewer

Contact: Dylan_Kotliar@hms.harvard.edu, Andres.Colubri@umassmed.edu

1. Introduction.

Dimensionality reduction methods such as UMAP (Becht et al. 2018) and tSNE (Amir et al. 2013) create 2D representations of scRNA-Seq data that preserve distances between cells, providing a visualization that captures much of the underlying data structure. scRNA-Seq analysis can be thought of as identifying, characterizing, and interpreting the biological signals that give rise to that structure. Software to aide in this task includes programmatic toolkits such as Scanpy (Wolf, Angerer, and Theis 2018) and SEURAT (Stuart et al. 2019) for Python and R respectively, and interactive viewers such as Single Cell Explorer (Feng et al. 2019) and CellXGene VIP (Li et al. 2020). While programmatic toolkits provide flexible commands for preprocessing, statistical analysis, and plotting of scRNA-Seq data, they interface with the user solely via programming commands and don't allow visual interaction with the embedding (e.g. selecting cells). Interactive interfaces enable direct visual interaction but generally do not support the flexibility of the programmatic toolkit. To our knowledge, no interactive scRNA-Seq visualization tool currently supports real-time transfer of data and results to and from the programming environment. Such transferability could enable users to rapidly iterate between interactive discovery of visual patterns, and computational analysis to validate those patterns. We therefore developed Sciviewer to facilitate interactive visual exploration of 2D embedding from within the Python programming environment.

2. Methods

Sciviewer is implemented with the Processing data visualization API in Java (https://processing.org/) which is accessible from within Python via the Py5 package (http://py5.ixora.io/about/). We leverage the hardware-accelerated rendering engine in Processing (Colubri and Fry, 2012), which can handle complex geometries in real time, to visualize large scRNA-seq datasets during interactive manipulation. It requires two inputs: (1) a gene expression matrix **X** (N cells X G genes, $X_{i,g}$ denotes expression of gene g in cell i), and any 2D embedding of the data such as UMAP - **E** (N cells X 2 dimensions, $(E_{i,x}, E_{i,y})$ denotes coordinates for cell$i$).

Sciviewer is launched from Python, and opens as a graphical interface that includes an interactive scatter plot of the embedding (Figure 1). Users can select a group of cells $\{i_1 \dots i_k\}$ to compute differential expression between selected and unselected cells. Sciviewer then shows the list of the most differentially expressed genes (defined via Welch's T-test), alongside violin plots of user-selected genes (Figure 1C). Alternatively, Sciviewer can identify genes that vary locally along any direction in the embedding (Figure 1B). Users select a set of cells and a direction$v = (v_x, v_y)$ and Sciviewer calculates the vector projection of the selected cells onto that direction and displays the genes with the greatest Pearson correlation (R_g) between the projected coordinates and gene expression. Mathematically, for gene$g$:

$$p_{i_j} = \frac{(E_{i_j,x}, E_{i_j,y}) \cdot v}{||v||} \quad j = 1 \dots k \quad ; \quad R_g = pearson(p_{i_1} \dots p_{i_k}, X_{i_1,g} \dots X_{i_k,g})$$

This is analogous to pseudotemporal ordering (Saelens et al. 2019), but the ordering is defined by a user-selected direction, allowing for rapid and flexible interrogation of the embedding. Notably, actions in Sciviewer cause real-time updates to corresponding variables in Python so

users can programmatically access the selected cells, and associated genes, test statistics, and P-values, for downstream programmatic analyses such as gene-set enrichment (Figure 1D).

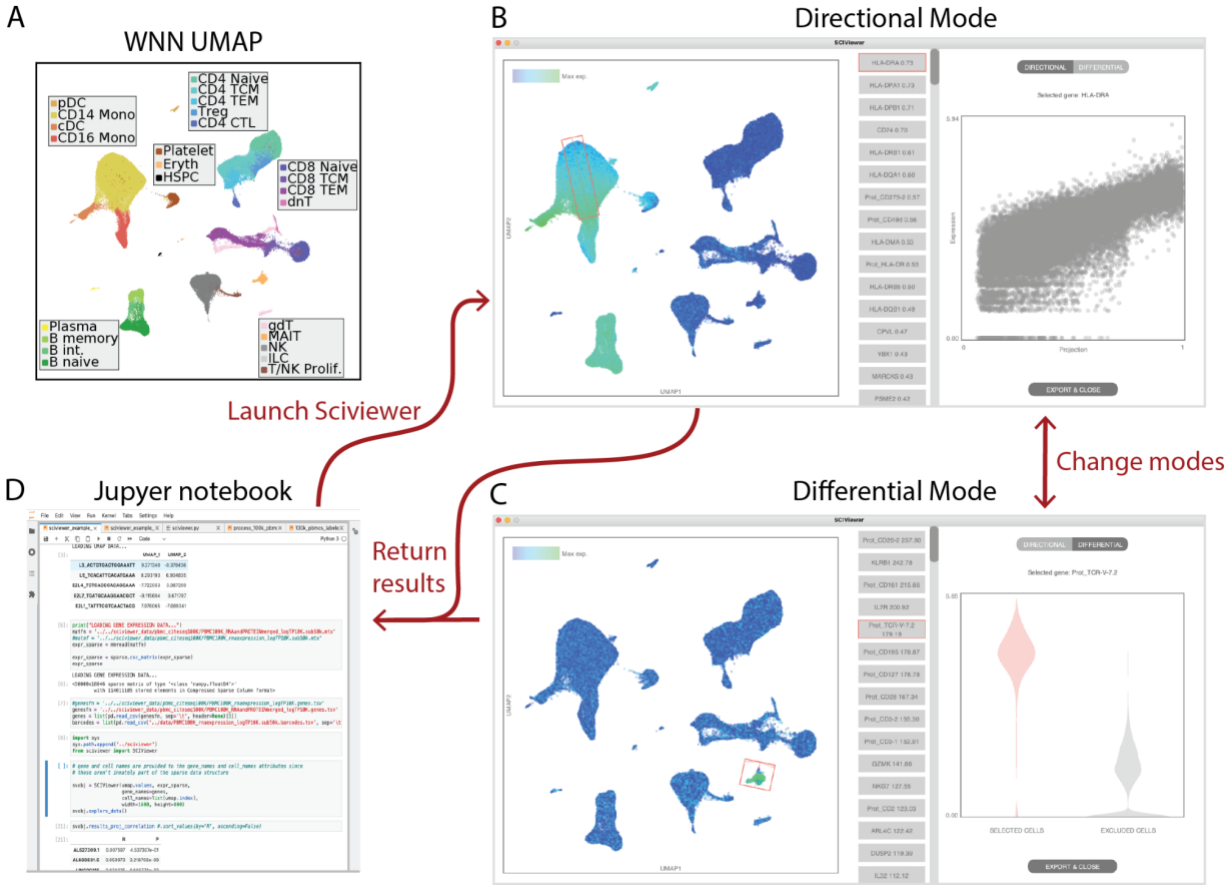


Figure 1. Application of Sciviewer to multimodal PBMC dataset

(**A**) UMAP embedding of CITE-Seq data of peripheral blood mononuclear cells (PBMCs) described in (Hao et al. 2021). Cells are labeled based on clusters described in that paper, with related cell-types aggregated for ease of visualization. (**B-C**) Screenshots of Sciviewer in directional and differential mode respectively for different user selections. (**D**) Screenshot of a Jupyter notebook environment, from which Sciviewer is called, and in which results of Sciviewer selections and calculations are available for programmatic analysis in real time.

3. Results

To illustrate the insights obtainable with Sciviewer, we applied it to a CITE-Seq dataset of 161,764 circulating immune cells, and 17,516 genes, consisting of transcriptome-wide profiling and targeted antibody-based capture of 211 proteins (Hao et al. 2021). We use Sciviewer to explore the novel weighted nearest neighbor-based UMAP described in the paper (Figure 1A), which intelligently weights protein and RNA data to generate the embedding. This demonstrates how Sciviewer is agnostic to the choice of 2D embedding and allows us to characterize signal from both RNA and protein features. Directional analysis of CD14+ monocytes demonstrated a gradient in expression of multiple HLA genes (responsible for antigen presentation) at both the RNA and protein levels, thus connecting a biological signal to the organization of the embedding (Figure 1B). Selecting a cluster of cells labeled as mucosal associated invariant T-cells (MAITs) in directional mode, we note a T-cell receptor V-segment protein that is not associated with any of the other T-cell populations, indicating the “invariant” receptor aspect of this T-cell population

(Figure 1C). On a 3.8 GHz 8-core Intel Core i7 Mac desktop computer, for this large dataset, it took 3.69 seconds to compute directional correlations for a selection of 25,531 cells, and 7.3 seconds to compute differential expression for 20,661 cells compared against 141,103 others, demonstrating the performance of the tool for a large dataset. This dataset and others are available as part of Sciviewer tutorials in the Github repository.

In summary, Sciviewer enables interactive exploration of scRNA-Seq that is tightly integrated with programmatic analysis in Python. It also introduces a novel directional association analysis that enables flexible exploration and interpretation of 2D embeddings. This approach could potentially have broad utility for other high-dimensional data types beyond scRNA-Seq.

Funding

The project described was supported by award Number T32GM007753 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

Amir, El-Ad David, Kara L. Davis, Michelle D. Tadmor, Erin F. Simonds, Jacob H. Levine, Sean C. Bendall, Daniel K. Shenfeld, Smita Krishnaswamy, Garry P. Nolan, and Dana Pe'er. 2013. "viSNE Enables Visualization of High Dimensional Single-Cell Data and Reveals Phenotypic Heterogeneity of Leukemia." *Nature Biotechnology* 31 (6): 545–52.

Becht, Etienne, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W. H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. 2018. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology*, December. https://doi.org/10.1038/nbt.4314.

Colubri, Andres and Fry, Ben. "Introducing processing 2.0." In *ACM SIGGRAPH 2012 Talks*, SIGGRAPH '12, pages 12:1--12:1, New York, NY, USA, 2012. ACM

Feng, Di, Charles E. Whitehurst, Dechao Shan, Jon D. Hill, and Yong G. Yue. 2019. "Single Cell Explorer, Collaboration-Driven Tools to Leverage Large-Scale Single Cell RNA-Seq Data." *BMC Genomics* 20 (1): 676.

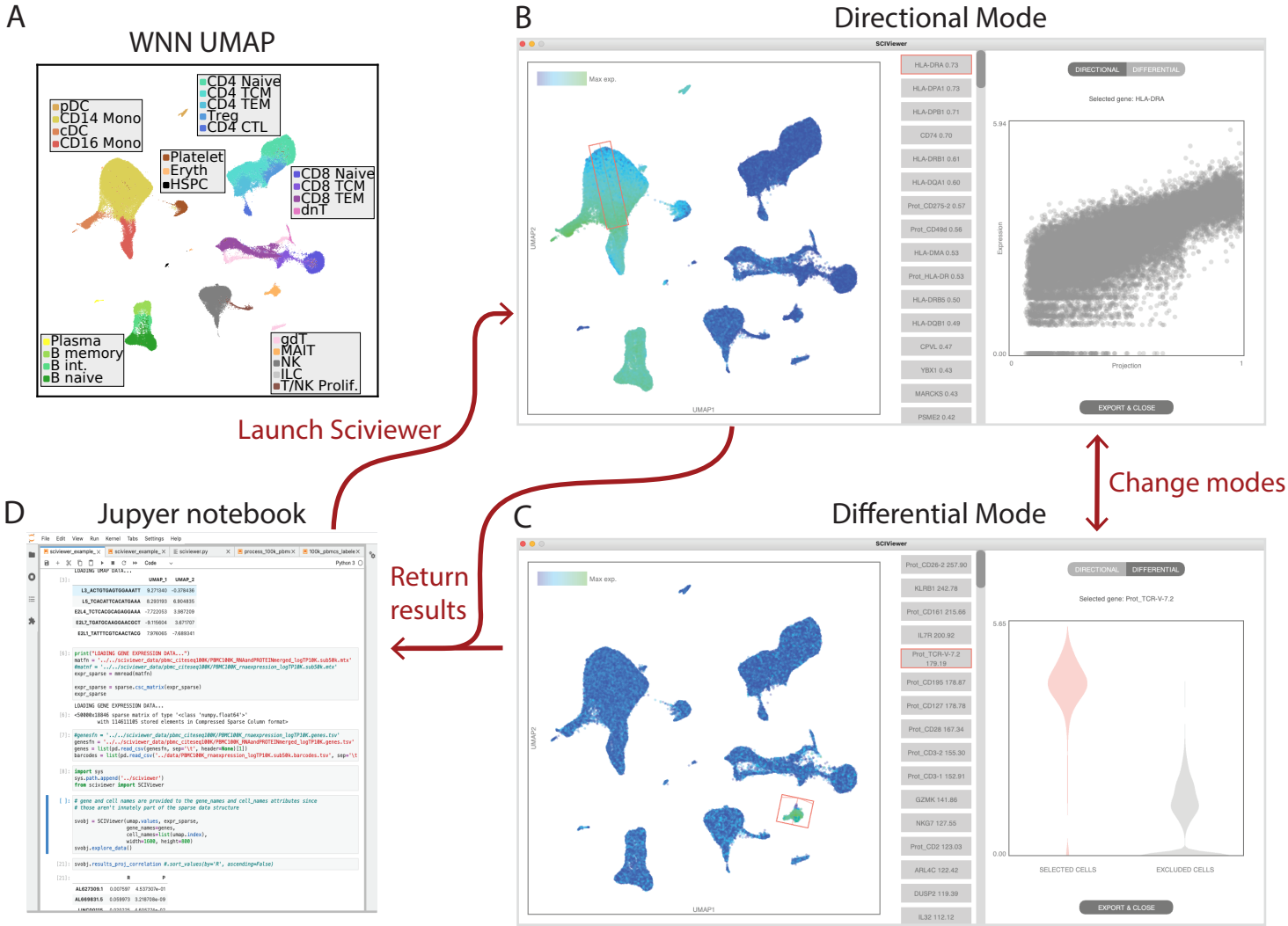
Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–87.e29.

Li, K., Z. Ouyang, D. Lin, M. Mingueneau, and W. Chen. 2020. "Cellxgene VIP Unleashes Full Power of Interactive Visualization, Plotting and Analysis of scRNA-Seq Data in the Scale of Millions of Cells." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.08.28.270652v1.abstract.

Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. 2019. "A Comparison of Single-Cell Trajectory Inference Methods." *Nature Biotechnology* 37 (5): 547–54.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888–1902.e21.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1): 15.


A
WNN UMAP
pDC
CD14 Mono
cDC
CD16 Mono
CD4 Naive
CD4 TCM
CD4 TEM
Treg
CD4 CTL
Platelet
Eryth
HSPC
CD8 Naive
CD8 TCM
CD8 TEM
dnT
Plasma
B memory
B int.
B naive
gdT
MAIT
NK
ILC
T/NK Prolif.
B
Directional Mode
SCIViewer
Max exp.
UMAP1
UMAP2
HLA-DRA 0.73
HLA-DPA1 0.73
HLA-DPB1 0.71
CD74 0.70
HLA-DRB1 0.61
HLA-DQA1 0.60
Prot_CD275-2 0.57
Prot_CD49d 0.56
HLA-DMA 0.53
Prot_HLA-DR 0.53
HLA-DRB5 0.50
HLA-DQB1 0.49
CPVL 0.47
YBX1 0.43
MARCKS 0.43
PSME2 0.42
DIRECTIONAL
DIFFERENTIAL
Selected gene: HLA-DRA
5.94
0.00
Expression
0
1
Projection
EXPORT & CLOSE
Launch Sciviewer
Change modes
D
Jupyer notebook
Return results
C
Differential Mode
SCIViewer
Max exp.
UMAP1
UMAP2
Prot_CD26-2 257.90
KLRB1 242.78
Prot_CD161 215.66
IL7R 200.92
Prot_TCR-V-7.2 179.19
Prot_CD195 178.87
Prot_CD127 178.78
Prot_CD28 167.34
Prot_CD3-2 155.30
Prot_CD3-1 152.91
GZMK 141.86
NKG7 127.55
Prot_CD2 123.03
ARL4C 122.42
DUSP2 119.39
IL32 112.12
DIRECTIONAL
DIFFERENTIAL
Selected gene: Prot_TCR-V-7.2
5.65
0.00
SELECTED CELLS
EXCLUDED CELLS
EXPORT & CLOSE