Bioinformatics, 37(21), 2021, 3961–3963
doi: 10.1093/bioinformatics/btab689
Advance Access Publication Date: 2 October 2021
Applications Note

Gene expression

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

Dylan Kotliar [1,2] **and Andrés Colubri**[2,3,*]

[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 and [3]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester MA, 01655, USA

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on July 18, 2021; revised on August 25, 2021; editorial decision on September 20, 2021; accepted on September 28, 2021

Abstract

Motivation: Visualizing two-dimensional embeddings (such as UMAP or tSNE) is a useful step in interrogating single-cell RNA sequencing (scRNA-Seq) data. Subsequently, users typically iterate between programmatic analyses (including 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 most of the data analysis takes place, interrupting the iterative process.

Results: 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: andres.colubri@umassmed.edu

1 Introduction

Dimensionality reduction methods such as UMAP (Becht *et al.*, 2018) and tSNE (Amir *et al.*, 2013) create two-dimensional (2D) representations of scRNA-Seq data that seek to preserve nearest neighbor relationships 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 aid in this task includes programmatic toolkits such as *Scanpy* (Wolf et al., 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 predominantly interface with the user via programming commands and do not enable sophisticated visual interaction with the embedding. While *SEURAT* provides the ability to select cells on an embedding via the CellSelector function, to our knowledge no corresponding functionality exists in *Scanpy*. Interactive viewers enable direct visual interaction but generally do not support the flexibility of the programmatic toolkit. We are not aware of any interactive scRNA-Seq viewer that 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 Single-cell Interactive Viewer (*Sciviewer*) to facilitate interactive visual exploration of 2D embedding from within the Python programming environment.

2 Materials and 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 (https://py5.ixora.io). 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: a gene expression matrix X (N cells by 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 by 2 dimensions, $(E_{i,x}, E_{i,y})$ denotes coordinates for cell i). Sciviewer supports sparse matrix formats for X, which substantially speeds up calculations.

Sciviewer is launched from Python and opens as a graphical interface that includes an interactive scatter plot of the embedding (Fig. 1). Users can select a group of cells $\{i_1 \ldots i_k\}$ 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 (Fig. 1C). Alternatively, *Sciviewer* can identify genes that vary locally along any direction in the embedding (Fig. 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 and cells j = 1 … k:

$$p_{i_j} = \frac{(E_{i_j,k}\ ,\ E_{i_j,k}) \cdot v}{||v||};\ R_g = \text{pearson}(p_{i_1} \ldots\ p_{i_k},\ X_{i_1,g} \ \ldots\ X_{i_k,g})$$

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, associated genes, test statistics, and *P*-values, for downstream programmatic analyses such as gene-set enrichment (Fig. 1D).

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 explore the novel weighted nearest neighbor-based UMAP described in the article (Fig. 1A), which intelligently weights protein and RNA modalities to generate the embedding. This demonstrates how *Sciviewer* is agnostic to the choice of 2D embedding and allows us to characterize signals from both RNA and protein features. Applied to the embedding and processed dataset, *as published*, 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 (Fig. 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, which underlies the 'invariant' receptor aspect of this T-cell population (Fig. 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

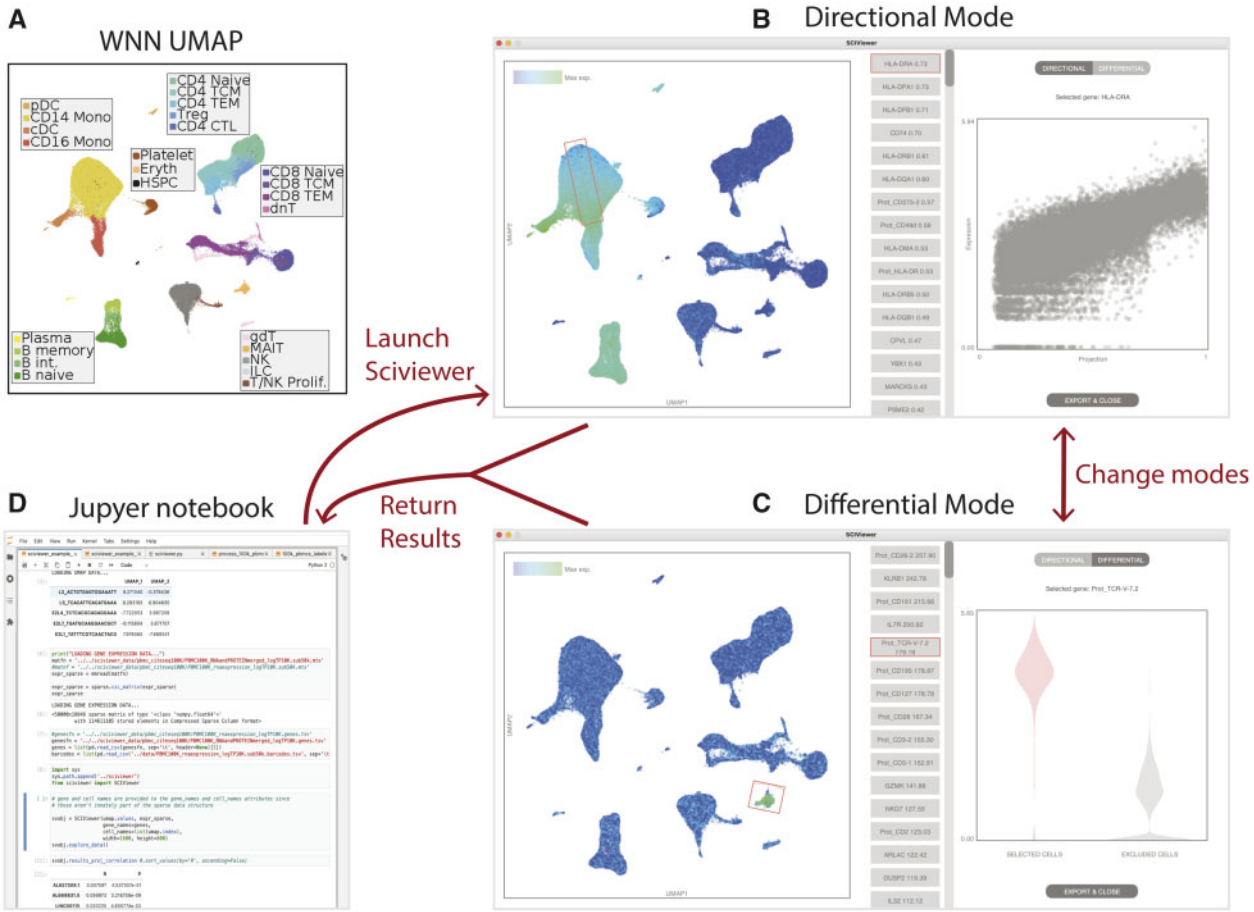

Fig. 1. Application of *Sciviewer* to multimodal PBMC dataset. (**A**) Weighted Nearest Neighbors (WNN) UMAP embedding of CITE-Seq data and 5' scRNA-Seq from peripheral blood mononuclear cells (PBMCs) described in Hao *et al.* (2021). Cells are labeled based on clusters described in that article, with 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.

dataset. This dataset and others are available as part of the *Sciviewer* tutorials in the Github repository.

In summary, *Sciviewer* enables interactive exploration of scRNA-Seq data 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.

Acknowledgement

The authors thank Jim Schmitz, the creator of the Py5 software that makes this work possible, and well as Pardis Sabeti and Aaron Lin for useful discussions and support.

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.

Conflict of Interest: none declared.

References

Amir,E.-A.D. *et al.* (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–552.

Becht,E. *et al.* (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, **37**, 38–44.

Colubri,A. and Fry,B. (2012) Introducing processing 2.0. In: *ACM SIGGRAPH 2012 Talks, SIGGRAPH '12*, 12:1. ACM, New York, NY, USA.

Feng,D. *et al.* (2019) Single cell explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data. *BMC Genomics*, **20**, 676.

Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.e29.

Li,K. *et al.* (2020) cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. *bioRxiv*.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.

Stuart,T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.e21.

Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.