SCelVis: exploratory single cell data analysis on the desktop and in the cloud

Benedikt Obermayer Equal first author, [1, 2], **Manuel Holtgrewe** Equal first author, [1, 2], **Mikko Nieminen** [1, 3], **Clemens Messerschmidt** [1, 2], **Dieter Beule** Corresp. [1, 3]

[1] Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany

[2] Charité – Universitätsmedizin Berlin, Berlin, Germany

[3] Max Delbrück Center for Molecular Medicine, Berlin, Germany

Corresponding Author: Dieter Beule
Email address: dieter.beule@bihealth.de

Background: Single cell omics technologies present unique opportunities for biomedical and life sciences from lab to clinic, but the high dimensional nature of such data poses challenges for computational analysis and interpretation. Furthermore, FAIR data management as well as data privacy and security become crucial when working with clinical data, especially in cross-institutional and translational settings. Existing solutions are either bound to the desktop of one researcher or come with dependencies on vendor-specific technology for cloud storage or user authentication.

Results: To facilitate analysis and interpretation of single-cell data by users without bioinformatics expertise, we present SCelVis, a flexible, interactive and user-friendly app for web-based visualization of pre-processed single-cell data. Users can survey multiple interactive visualizations of their single cell expression data and cell annotation, define cell groups by filtering or manual selection and perform differential gene expression, and download raw or processed data for further offline analysis. SCelVis can be run both on the desktop and cloud systems, accepts input from local and various remote sources using standard and open protocols, and allows for hosting data in the cloud and locally. We test and validate our visualization using publicly available scRNA-seq data.

Methods: SCelVis is implemented in Python using Dash by Plotly. It is available as a standalone application as a Python package, via Conda/Bioconda and as a Docker image. All components are available as open source under the permissive MIT license and are based on open standards and interfaces, enabling further development and integration with third party pipelines and analysis components. The GitHub repository is https://github.com/bihealth/scelvis.

SCelVis: exploratory single cell data analysis on the desktop and in the cloud

Benedikt Obermayer[1,2,*], Manuel Holtgrewe[1,2,*], Mikko Nieminen[1,3], Clemens Messerschmidt[1,2], Dieter Beule[1,3]

[1] Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany
[2] Charité – Universitätsmedizin Berlin, Berlin, Germany
[3] Max Delbrück Center for Molecular Medicine, Berlin, Germany

* these authors contributed equally

Corresponding Author:
Dieter Beule[1,3]
Chariteplatz 1, 10117 Berlin, Germany
Email address: dieter.beule@bihealth.de

Abstract

Background: Single cell omics technologies present unique opportunities for biomedical and life sciences from lab to clinic, but the high dimensional nature of such data poses challenges for computational analysis and interpretation. Furthermore, FAIR data management as well as data privacy and security become crucial when working with clinical data, especially in cross-institutional and translational settings. Existing solutions are either bound to the desktop of one researcher or come with dependencies on vendor-specific technology for cloud storage or user authentication.

Results: To facilitate analysis and interpretation of single-cell data by users without bioinformatics expertise, we present SCelVis, a flexible, interactive and user-friendly app for web-based visualization of pre-processed single-cell data. Users can survey multiple interactive visualizations of their single cell expression data and cell annotation, define cell groups by filtering or manual selection and perform differential gene expression, and download raw or processed data for further offline analysis. SCelVis can be run both on the desktop and cloud systems, accepts input from local and various remote sources using standard and open protocols, and allows for hosting data in the cloud and locally. We test and validate our visualization using publicly available scRNA-seq data.

Methods: SCelVis is implemented in Python using Dash by Plotly. It is available as a standalone application as a Python package, via Conda/Bioconda and as a Docker image. All components are available as open source under the permissive MIT license and are based on open standards and interfaces, enabling further development and integration with third party pipelines and analysis components. The GitHub repository is https://github.com/bihealth/scelvis.

Introduction

Single-cell omics technologies, in particular single-cell RNA sequencing (scRNA-seq), allow for the high-throughput profiling of gene expression in thousands to millions of cells with unprecedented resolution. Recent large-scale efforts are underway to catalogue and describe all human cell types (Regev et al., 2017) and to study cells and tissues in health and disease (https://lifetime-fetflagship.eu). Single-cell sequencing could therefore become a routine tool in the clinic for comprehensive assessments of molecular and physiological alterations in diseased organs as well as systemic responses, e.g., of the immune system. The enormous scale and high-dimensional nature of the resulting data presents an ongoing challenge for computational analysis (Stegle, Teichmann & Marioni, 2015). Ever more sophisticated methods, e.g., deep learning frameworks (Eraslan et al., 2019), extract multiple layers of information from cell types to lineages and differentiation programs. Many of these methods, their mathematical background, and the underlying assumptions will remain opaque to users without specific bioinformatics expertise. At the same time, an in-depth understanding of the relevant biology is often beyond the know-how of typical bioinformatics researchers. More than ever, single-cell omics requires close communication and collaboration from wet and dry lab experts. Due to the large amount of data, communication needs to be based on interactive channels (e.g., web-based apps) rather than static tables. Further, as single-cell omics moves towards the clinic, FAIR (Wilkinson et al., 2016) data management, data privacy, and data security issues need to be handled appropriately. All employed methods should be able to scale towards handling a large number of users and even larger numbers of samples.

State of the Art. Web apps have been used extensively in the single-cell literature and are most commonly built on Shiny (RStudio Inc., 2014). However, standalone and general-purpose tools are still quite rare. Pagoda (Fan et al., 2016) comes with a simple intuitive web app, which is limited to Pagoda output and requires manual preprocessing. Cerebro (Hillje, Pelicci & Luzi, 2019) is a Shiny web app and provides relatively rich functionality such as gene set enrichments and quality control statistics, but requires extensive manual preprocessing and is not (yet) ready for larger frameworks. The Single Cell Viewer (SCV; Wang et al., 2019) also relies on Shiny, but its input is limited to Seurat objects. CellexalVR (Legetth et al., 2018) provides an immersive virtual reality platform for the visualization and analysis of scRNA-seq data, but requires special hardware. Cellxgene (https://chanzuckerberg.github.io/cellxgene/) is very fast and user-friendly but restricted to visualizing two-dimensional embeddings. Finally, the Broad Single Cell Portal (https://portals.broadinstitute.org/single_cell) provides a large-scale web service for a large number of users and studies. It includes a 10X Genomics data processing pipeline and user authentication/account management. However, the underlying Docker image strongly depends on vendor-specific cloud systems such as Google cloud and Broad Firecloud services. Its implementation thus poses practical hurdles, in particular if it is to be integrated into existing clinical infrastructure.

Materials & Methods

SCelVis is based on Dash by Plotly (Plotly Technologies Inc., 2015) and accepts data in HDF5 format as AnnData objects. These objects can be created using Scanpy (Wolf, Angerer & Theis, 2018), provide a scalable and memory-efficient data format for scRNA-seq data and integrate naturally into python environments. SCelVis also provides conversion functionality to AnnData from raw text, loom format or 10X Genomics CellRanger output. The built-in converter is accessible from the command line and a web-based user interface (Figure 1). One HDF5 file or a folder containing multiple such files can then be provided to SCelVis for visualization, and data sets can be selected for exploration on the graphical web interface. To enable both local and cloud access, data can be read from the file system or remote data sources via the standard internet protocols FTP, SFTP, and HTTP(S). SCelVis also provides data access through the open source iRODS protocol (Rajasekar et al., 2010) or the widely-used Amazon S3 object storage protocol. The data sources can be given on the command line and as environment variables as is best practice for cloud deployments (Adam Wiggins, 2011). The latter allows for easy "serverless" and cloud deployments.

SCelVis is built around two viewpoints on single-cell data (Figure 1). On the one hand, it provides a cell-based view, where users can browse and investigate cell annotations (e.g., cell type) and cell-specific statistics (sequencing depth, cell type proportions, etc.) in multiple visualizations, e.g., on a t-SNE or UMAP embedding, as violin plots or bar charts. Cells to be displayed can be filtered by various criteria, and groups of cells can be defined manually on a scatter plot as input for on-the-fly differential gene expression analysis. On the other hand, SCelVis provides a gene-based view that lets users explore gene expression in multiple visualizations on embeddings or as violin or dot plots. Relevant genes can be specified by hand or selected directly from lists of marker or differential genes.

The source code is available under the permissive MIT license on the GitHub repository at https://github.com/bihealth/scelvis, which also contains a tutorial movie and a link to a public demonstration instance. The software can be run both in the cloud and on workstation desktops via Docker.

Usage Example

We provide three example datasets within our Github repository or via figshare. First, a small synthetic simulated dataset created for testing and illustration purposes, and secondly a publicly available processed scRNA-seq dataset from 10X Genomics containing ~1000 cells of a mix of human HEK293T and murine NIH3T3 cells. Finally, we reanalyzed a published data set of stimulated and control peripheral blood mononuclear cells (PBMSc; Kang et al., 2018) with the Seurat "data integration" workflow (Stuart et al., 2019) and made it accessible via https on figshare (https://files.figshare.com/18037739/pbmc.h5ad; Figure 2). With the species-mix dataset from 10X, the relevant plot to demonstrate a low doublet rate can be readily re-created (Fig. 2A left; compare to Fig. 2a in Zheng et al. (2017), which shows data obtained with a previous version of the 10X chemistry), and the species composition of the different clusters found by CellRanger can be easily interrogated (Fig. 2A right). For the PBMC dataset, it is straightforward to perform differential gene expression analysis, e.g., between stimulated and control monocytes by using the "filter" and "differential gene expression" buttons (Fig. 2B). Summarized gene expression for cell-type marker genes as well as for general (e.g., IFI6) or cell-type specific (e.g., CXCL10) differential genes can be displayed in a split dot plot as in Fig. 2d of Stuart, Butler et al. 2019. Hence, our visualizations for the published datasets are equivalent to those obtained from other visualization tools, e.g., Seurat.

Conclusions

In this manuscript, we have presented SCelVis, a method for the interactive visualization of single-cell RNA-seq data. It provides easy-to-use yet flexible means of scRNA-seq data exploration for researchers without computational background. SCelVis takes processed data, e.g., provided by CellRanger or a bioinformatics collaboration partner, as input, and focuses solely on visualization and explorative analysis. Great care has been taken to make the method flexible in usage and deployment. It can be used both on a researcher's desktop with minimal training yet its usage scales up to a cloud deployment. Data can be read from local file systems but also from a variety of remote data sources, e.g., via the widely deployed (S)FTP, S3, and HTTP(S) protocols. This allows for deploying it in a Docker container on "serverless" cloud systems. As both the application and data can be hosted on the network or cloud systems, the application facilitates cross-institutional research. For example, a sequencing or bioinformatics core unit can use it for giving access to non-computational collaboration partners over the internet. This is particularly relevant as it comes with no dependency on any vendor-specific technology such as the Google or Facebook authentication that appears to become pervasive in today's life science.

Acknowledgements

The example dataset for the 1:1 mixture of human and mouse cells processed with CellRanger (v3) was taken from the 10X genomics website https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/hgmm_1k_v3. The example dataset for the stimulated vs. control PBMCs was taken from GEO (accession GSE96583) and re-analyzed with the Seurat sample alignment strategy as explained in the tutorial at https://satijalab.org/seurat/v2.4/immune_alignment.html.

References

Adam Wiggins. 2011.The Twelve-Factor App. *Available at https://12factor.net/* (accessed March 22, 2019).

Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* 20:389–403. DOI: 10.1038/s41576-019-0122-6.

Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J, Kharchenko P V. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods* 13:241–244. DOI: 10.1038/nmeth.3734.

Hillje R, Pelicci PG, Luzi L. 2019. Cerebro: Interactive visualization of scRNA-seq data. *bioRxiv*:631705. DOI: 10.1101/631705.

Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* 36:89–94. DOI: 10.1038/nbt.4042.

Legetth O, Rodhe J, Pålsson J, Wallergård M, Lang S, Soneji S. 2018. CellexalVR: A virtual reality platform for the exploration and analysis of single-cell gene expression data. *bioRxiv*:329102. DOI: 10.1101/329102.

Plotly Technologies Inc. 2015. Collaborative data science.

Rajasekar A, Moore R, Hou C-Y, Lee CA, Marciano R, de Torcy A, Wan M, Schroeder W, Chen S-Y, Gilbert L, Tooby P, Zhu B. 2010. iRODS Primer: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2:1–143. DOI: 10.2200/S00233ED1V01Y200912ICR012.

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N, Human Cell Atlas Meeting Participants. 2017. The Human Cell Atlas. *eLife* 6. DOI: 10.7554/eLife.27041.

RStudio Inc. 2014. shiny: Easy web applications in R.

Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 16:133–145. DOI: 10.1038/nrg3833.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* 177:1888-1902.e21. DOI: 10.1016/j.cell.2019.05.031.

Wang S, Brett C, Bolisetty M, Golhar R, Neuhaus I, Ravi K. 2019. Single Cell Viewer (SCV): An interactive visualization data portal for single cell RNA sequence data. *bioRxiv*:664789. DOI: 10.1101/664789.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Santos LB da S, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PA. 't, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Schaik R van, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Lei J van der, Mulligen E van, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3*.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* 19:15. DOI: 10.1186/s13059-017-1382-0.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 8. DOI: 10.1038/ncomms14049.

Figure 1

Overview of SCelVis Architecture and User Interface

A: Data can be converted from CellRanger output, loom format or raw text to an input HDF5 file with the SCelVis converter. These files can be uploaded into the web app or loaded remotely via various protocols such as S3, HTTP, etc. SCelVis can then be run locally or on a server/in the cloud and provides various views of the analysis results. **B:** screenshot of the SCelVis interface for a mixture of human and mouse cells from 10X genomics. Users can browse the "about" tab to obtain background information on the data (1), select the "cell annotation" tab (2) to investigate cell meta data or the "gene expression" tab (3) to interrogate gene expression. The cell annotation view provides scatter, violin and bar plots (4). Displayed cells can be filtered (5) by a number of criteria. In typical cases, the scatter plot would be configured with embedding variables on the x- and y-axis (6) and a categorical or continuous variable for the coloring (7). Differential gene expression (8) can be performed by manually selecting groups of cells on the scatter plot, using "box select" or "lasso select" in hover bar on the top right-hand corner of the plot (9). Here, plot results can also be downloaded in png format. The underlying data can be obtained from a link at the bottom left (10). Other datasets can be selected, uploaded or converted from the menu on the top right (11).

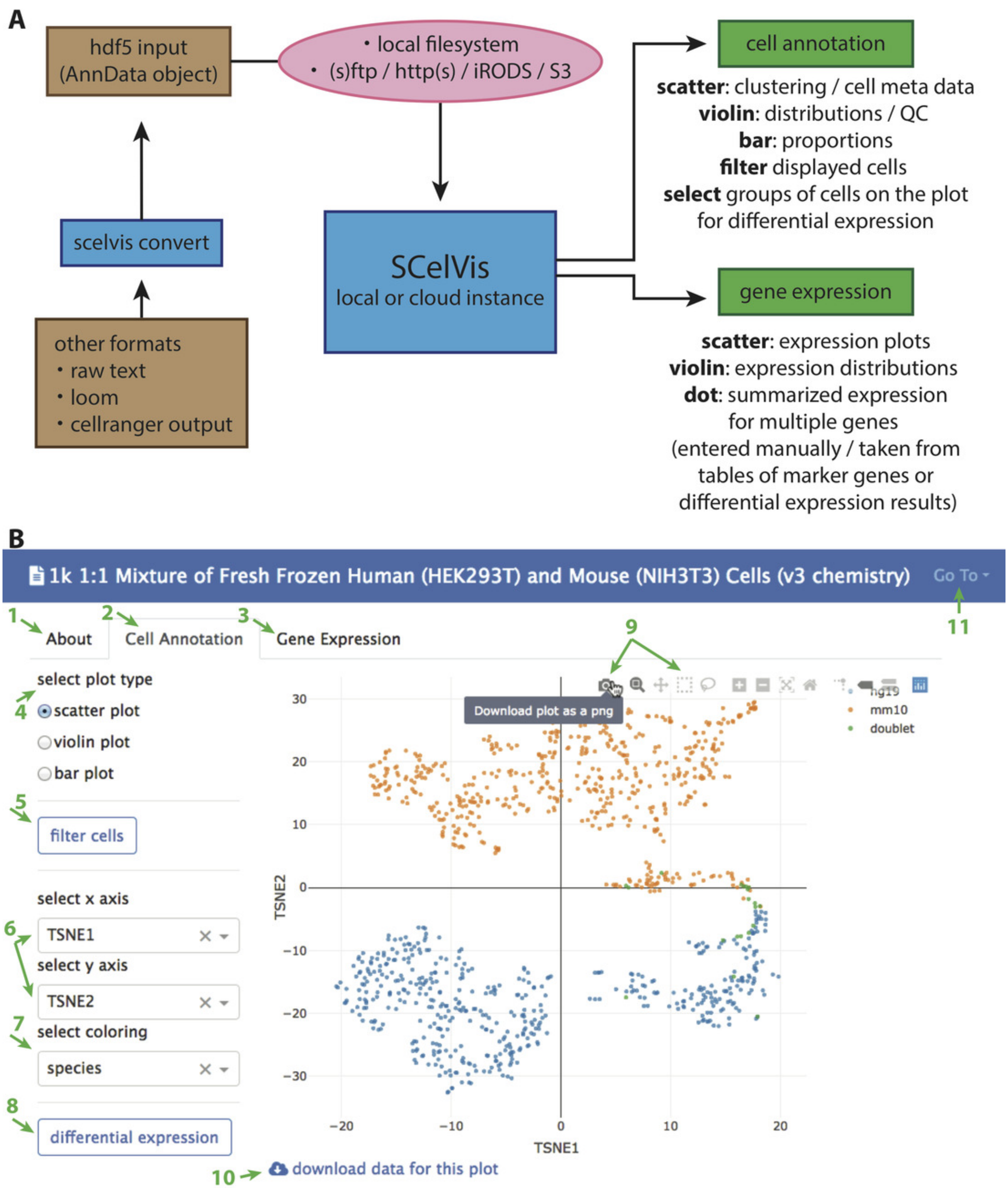
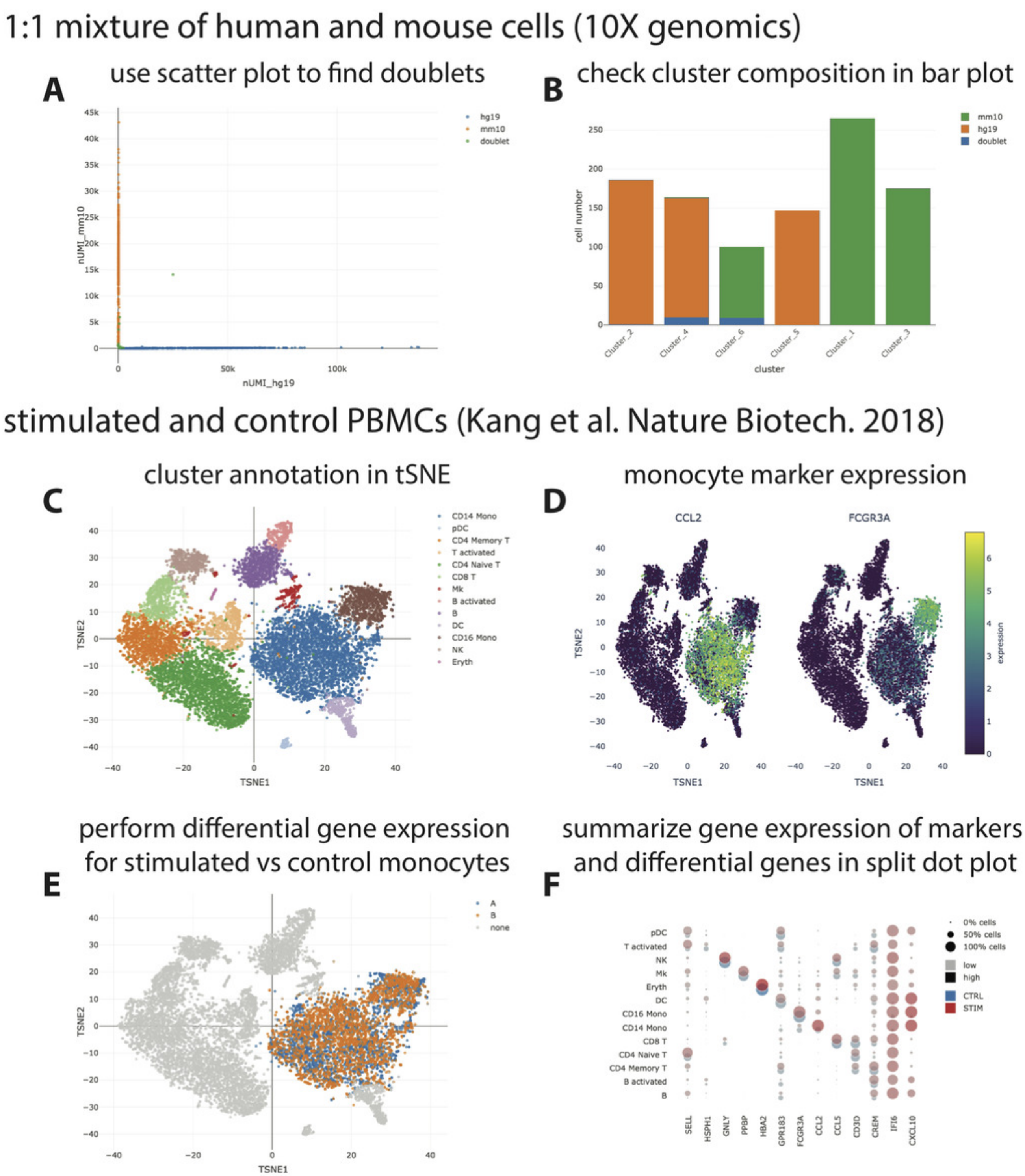

A
hdf5 input (AnnData object)
• local filesystem
• (s)ftp / http(s) / iRODS / S3
cell annotation
scatter: clustering / cell meta data
violin: distributions / QC
bar: proportions
filter displayed cells
select groups of cells on the plot for differential expression
scelvis convert
SCelVis
local or cloud instance
gene expression
scatter: expression plots
violin: expression distributions
dot: summarized expression for multiple genes (entered manually / taken from tables of marker genes or differential expression results)
other formats
• raw text
• loom
• cellranger output
B
1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells (v3 chemistry)
Go To
About
Cell Annotation
Gene Expression
select plot type
scatter plot
violin plot
bar plot
filter cells
select x axis
TSNE1
select y axis
TSNE2
select coloring
species
differential expression
Download plot as a png
hg19
mm10
doublet
TSNE1
TSNE2
download data for this plot
1
2
3
4
5
6
7
8
9
10
11

Figure 2

Visualization of publicly available scRNA-seq data.

A+B: scRNA-seq data for a 1:1 mixture of 1k fresh frozen human (HEK293T) and mouse (NIH3T3) cells (Chromium v3 chemistry) were taken from the 10X website (CellRanger output) and visualized with SCelVis. A scatter plot shows human vs. mouse UMI counts per cell and confirms a low doublet rate **(A)**, while a bar plot visualizes the species composition of the different clusters defined by CellRanger **(B)**. **C-F:** scRNA-seq data for stimulated vs. control PBMCs (Kang et al. 2018). The cluster annotation resulting from the Seurat sample alignment workflow (https://satijalab.org/seurat/v2.4/immune_alignment.html) can be interrogated and monocyte markers can be displayed by selecting from a table of marker genes (**C+D**). Stimulated or control monocytes can then be isolated using "filter cells" and defined as groups "A" or "B", respectively, for differential expression analysis (**E**). Summarized gene expression can be displayed for marker genes as well as cell-type specific or globally differential genes in a split dot plot (**F**).


1:1 mixture of human and mouse cells (10X genomics)
A
use scatter plot to find doublets
hg19
mm10
doublet
nUMI_mm10
nUMI_hg19
B
check cluster composition in bar plot
mm10
hg19
doublet
cell number
cluster
Cluster_2
Cluster_4
Cluster_6
Cluster_5
Cluster_1
Cluster_3
stimulated and control PBMCs (Kang et al. Nature Biotech. 2018)
C
cluster annotation in tSNE
CD14 Mono
pDC
CD4 Memory T
T activated
CD4 Naive T
CD8 T
Mk
B activated
B
DC
CD16 Mono
NK
Eryth
TSNE1
TSNE2
D
monocyte marker expression
CCL2
FCGR3A
expression
TSNE1
TSNE2
E
perform differential gene expression
for stimulated vs control monocytes
A
B
none
TSNE1
TSNE2
F
summarize gene expression of markers
and differential genes in split dot plot
0% cells
50% cells
100% cells
low
high
CTRL
STIM
SELL
HSPH1
GNLY
PPBP
HBA2
GPR183
FCGR3A
CCL2
CCL5
CD3D
CREM
IFI6
CXCL10