SCHNAPPs - Single Cell sHiNy APPlication(s)

Bernd Jagla, Valentina Libri, Claudia Chica, Vincent Rouilly, Sebastien Mella, Michel Puceat, Milena Hasan

Journal of Immunological Methods 499 (2021) 113176

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

SCHNAPPs - Single Cell sHiNy APPlication(s)

Bernd Jagla [a,b,*], Valentina Libri [a], Claudia Chica [b], Vincent Rouilly [c], Sebastien Mella [a,b], Michel Puceat [d], Milena Hasan [a]

[a] Institut Pasteur, Université de Paris, Cytometry and Biomarkers UTechS, F-75015 Paris, France
[b] Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[c] Datactix, 40 rue Neuve, 33000, Bordeaux, France
[d] Aix-Marseille University, INSERM U-1251, MMG, France

ARTICLE INFO

Keywords:
scRNA-seq
multi-omics data analysis
CITE-Seq
Shiny application

ABSTRACT

Single-cell RNA-sequencing (scRNAseq) experiments are becoming a standard tool for bench-scientists to explore the cellular diversity present in all tissues. Data produced by scRNAseq is technically complex and requires analytical workflows that are an active field of bioinformatics research, whereas a wealth of biological background knowledge is needed to guide the investigation. Thus, there is an increasing need to develop applications geared towards bench-scientists to help them abstract the technical challenges of the analysis so that they can focus on the science at play. It is also expected that such applications should support closer collaboration between bioinformaticians and bench-scientists by providing reproducible science tools.

We present SCHNAPPs, a Graphical User Interface (GUI), designed to enable bench-scientists to autonomously explore and interpret scRNAseq data and associated annotations. The R/Shiny-based application allows following different steps of scRNAseq analysis workflows from Seurat or Scran packages: performing quality control on cells and genes, normalizing the expression matrix, integrating different samples, dimension reduction, clustering, and differential gene expression analysis. Visualization tools for exploring each step of the process include violin plots, 2D projections, Box-plots, alluvial plots, and histograms. An R-markdown report can be generated that tracks modifications and selected visualizations. The modular design of the tool allows it to easily integrate new visualizations and analyses by bioinformaticians. We illustrate the main features of the tool by applying it to the characterization of T cells in a scRNAseq and Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) experiment of two healthy individuals.

1. Introduction

Successful and efficient data analysis of single-cell experiments requires comprehensive knowledge of bioinformatics and biology and thus often involves close interaction between bioinformaticians and bench-scientists. While the first feed the data through an analysis pipeline, the second interpret the results. Given the iterative nature of analyses, these interactions are usually frequent. In general, a pipeline must be rerun multiple times to remove/select cells or genes from the analysis. For example, damaged cells with highly expressing mitochondrial and ribosomal genes should be removed. As well, genes expressed by irrelevant cell types or belonging to biological processes that are not relevant to the scientific question could be left out from the analysis. When comparing or visualizing specific cell types only a subset of cells can be used.

Many tools are being developed to tackle this challenge (a comprehensive selection is listed here: https://github.com/federicomarini/awesome-expression-browser), some of which have been covered in a recent review (Cakir et al., 2020). Among the most accomplished are iSEE (Rue-Albrecht et al., 2018), Cerebro (Hillje et al., 2020), ASAP (Gardeux et al., 2017), iS-CellR (Patel, 2018), and singleCellTK (Jenkins et al., 2018). Recently, Hao et al. (Hao et al., 2021) made Azimuth available, an extensible web-based system that closely follows the Seurat-pipeline with the possibility of adding further functionality. It is also worth

; scRNAseq, Single-cell RNA-sequencing; CITE-seq, Cellular Indexing of Transcriptomes and Epitopes by Sequencing; GUI, graphical user interface; ADT, Antibody Derived Tag; CSV, comma-separated values; TSV, tabulator-separated values; PBMC, peripheral blood mononuclear cell; pDC, plasmacytoid dendritic cell; GEO, Gene Expression Omnibus.

* Corresponding author at: Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, F-75015 Paris, France.
E-mail address: bernd.jagla@pasteur.fr (B. Jagla).

https://doi.org/10.1016/j.jim.2021.113176
Received 21 August 2020; Received in revised form 21 October 2021; Accepted 25 October 2021
Available online 4 November 2021

mentioning the SeuratV3Wizard (https://github.com/nasqar/seuratv3wizard), part of the NASQAR system (Yousif et al., 2020), which is also built around the Seurat pipeline. Since all these systems are developing rapidly and features are continuously added, a comparison of abilities that holds over time is not possible.

Here, we present SCHNAPPs (Single Cell sHiNy APPlication(s)), a R/Shiny application that aims at enabling bench-scientists to characterize individual cells and genes starting from the initial normalization steps of raw counts to differential expression analysis. The selection process is captured in a report that can be used by a bioinformatician to validate and optimize the results. The software architecture of the application makes it easy for bioinformaticians to integrate new visualizations and analyses. As such, SCHNAPPs streamlines the interaction of bench-scientists and bioinformaticians, thus speeding up analyses.

A key concept of SCHNAPPs is that all single attributes of a cell (other than the count for a given gene) are treated equally. SCHNAPPs regards all properties that can be related to a single cell as a "projection". Thus, tSNE (Donaldson, 2016) or UMAP (McInnes et al., 2020) coordinates, principal components, mitochondrial content, or expression of proteins from CITE-seq experiments are all considered equally projections as they can be all used to project the cells in a low dimensional space (2D or 3D). Cell-type assignments are added as new projections as well as time assignments derived from a trajectory inference. The user can use these projections as x/y/z coordinates or to color cells.

The different views in SCHNAPPs are "intelligent" in the sense that they render different types of graphs depending on the type of data shown (factorial, numerical, logical, cell identifiers). These concepts and features distinguish SCHNAPPs from most other tools. Another feature that we explore here in more depth is the possibility to work with multiple samples. SCHNAPPs has a unique combination of features that allows optimizing parameters for dimension reduction and clustering; visualization tools that allow to identify and characterize differential and common marker sets for manually selected cells and is extendable/adaptable to the changing needs of researchers.

SCHNAPPs should be regarded as an exploratory tool that allows bench-scientists to perform independent analyses including testing different parameters like, for example, different constellations of cells and genes. However, we strongly advise to validate any statistical findings with an expert.

2. Material and methods

Samples. The PBMC were extracted from the blood of two healthy individuals by gradient centrifugation. All participants gave written informed consent in the frame of the healthy volunteers CoSImmGEn cohort (Clinical trials NCT 03925272), after approval of the CPP Ile-de-France I Ethics Committee (2011, jan 18th)).

PBMC were either stained with CITE-seq antibodies (Total-seq B antibodies by Biolegend see Supplementary Table 1) or sorted by FACS. We used 3′ single-cell gene expression V3.1 kit (10× Genomics) and Total-seq B antibodies (Biolegend), according to standard protocols.

For the PBMC CITE-seq staining: 5 μl Human TruStain FcX were added to 1×10^6 of PBMC resuspended in 100 μl of Cell Staining Buffer (BioLegend, Cat# 420201) and incubated 10 min at 4 °C. 1uL of each antibody-oligonucleotide conjugate was used to prepare the antibody mix. The mix was centrifuged at 14,000 rcf for 10 min at 4 °C and the supernatant added to the cells and incubated 30 min at 4 °C in the dark. Cells were then washed 3 times with PBS/0.04% BSA.

For the FACS sorted T cells: we used anti-CD56 BV510 (Cat# 318339 Biolegend), CD14 APC-Cy7 (Cat# 325620 Biolegend) and DAPI (Cat# 62248 Thermo Fisher) to exclude NK, monocytes and dead cells, respectively. We used anti-CD8a PerCP (Cat# 344708 Biolegend) and anti-CD4 FITC (Cat# 344604 Biolegend) to positively select CD4 and CD8 T cells.

Both PBMC and sorted T cells were resuspended at 1×10^6/mL in PBS/0.04% BSA and 6000 cells were loaded on 10× Chromium controller using the Chromium NextGEM Single Cell 3' Reagent Kit v3.1 (CG000206_ChromiumNextGEMSingleCell3'v3.1_CellSurfaceProtein_Rev D), assuming a recovery rate of 50%. GEM Generation & Barcoding, Post GEM-RT Cleanup & cDNA Amplification, and 3' Gene Expression Library Construction were performed as per manufacturer's instructions. Libraries were mixed prior to sequencing on Illumina Novaseq 150PE at minimum 20 k paired reads per cell for the gene expression libraries and at 5 k paired reads per cell for the Antibody Derived Tag (ADT) libraries.

Cellranger 4.0.0 was used with GRCh38-2020-A as reference to calculate count matrices. Filtered feature bc matrices were used as input to the script dataPrep.R (supplementary materials).

3. Results

3.1. Implementation

Input to the application is either a simple count matrix of comma-separated values (CSV), with rows representing features/genes and columns representing cells, or a SingleCellExperiment object (Amezquita et al., 2019) with a sparse matrix holding the counts and annotations for the cells (covariates) and annotation data for the features/genes. The singleCellExperiment object must have the following gene information for each gene: "symbol", the gene-symbol; "id", a potentially different unique identifier; "Description", descriptive information for the gene. Cell-specific information must include "sampleNames", a string/factor to distinguish cells from different samples; and "barcode", a unique barcode per sample. In practice, additional gene-specific annotations like functional annotations from Ensembl (Yates et al., 2020), or cell-specific annotations like cell type predictions using from SingleR (Aran et al., 2019) or scLearn (Duan et al., 2020) are computed during data preparation on the command line and integrated into the singleCellExperiment object (see supplementary material, dataPrep.R). Generally, we encourage calculating all relevant features that are not dependent on the ensemble of cells/genes to be pre-computed and integrated beforehand. (Cell phase, cell type prediction, mitochondrial content).

Examples for generating these objects are given on GitHub (https://github.com/baj12/SCHNAPPsContributions#prepare-data-for-schnapps).

The processed data can be exported as singleCellExperiment objects. This allows creating a SingleCellExperiment object from a CSV file that is usable with iSEE or other tools based on the SingleCellExperiment object. Multiple SingleCellExperiment files can be loaded and analyzed together.

Reproducibility is achieved by the creation of a directory (the name of the directory contains the date and time SCHNAPPs was launched) that holds an R-markdown file (Xie et al., 2019) with associated data archiving all major data manipulations (removal of data, normalization, clustering) and plots that were saved from within the application. Thus, it is possible to reproduce the cell selection, validate the analysis steps and optimize the graph for final publication.

The shiny framework (Chang et al., 2021) is used as the underlying framework with the dashboard design (Chang and Borges Ribeiro, 2018) for the graphical user interface (GUI). It can be run from within RStudio (RStudio Team, 2020) or as a stand-alone web application (docker, virtual machine).

Internally, the SingleCellExperiment object is used to store count matrices and user-supplied annotation. To take full advantage of the reactive concept with its dependency graph, individual computations (normalizations, projections/covariates) are stored in distinct objects. This approach avoids recalculating objects that do not depend on parameters that have changed. Due to the low coverage and dropouts associated with most single-cell sequencing experiments, sparse matrices are used to represent raw and normalized read counts, reducing the memory footprint. Parallel implementations are used when

available, such as for tSNE, UMAP. Shiny modules allow reuse and standardization of visualizations. Violin plots, 2D plots, and tables are modularized and can be used by any other contributed functionality. The 2D plot module, for example, allows to select cells, re/define/name groups of cells, log transform data, or normalize it by e.g., gene count per cell, just for the given plot. Selected cell names can optionally be shown and thus copied and pasted. This provides the ability to refine the analysis by e.g., sub-clustering a set of cells within a given cluster and represents an important tool for identifying the phenotype and potential fate of cells.

Contributions allow adding analyses or visualization tools; the end-user provides the directory where the contributions are located on the file system during the startup of the application. The application then looks for specific file names that contain sources for the GUI elements and reactive objects. Contributions for trajectory inference (SCORPIUS (Cannoodt et al., 2016), ElPiGraph.R (Albergante, 2021), Tempora (Tran and Bader, 2019)), for imputation (DCA (Eraslan et al., 2019), and per-cell gene signatures (Cell-ID (Cortal et al., 2021)) are already available. A dummy contribution is available that holds example code for key features (adding projections, normalizations/imputations, visualizations, and reports), which serves as an entry point for developers. New normalization and imputation methods can be integrated as well as differential expression methods. This concept allows restricting the functionality to only those tools that are useful for a given biological question, thus reducing the complexity of the application.

A computer with substantial memory and CPUs is recommended for the use of SCHNAPPs (e.g., 64GB RAM allows working with ~80,000 cells and ~ 20,000 genes).

3.2. Example use case

Single-cell RNA-Seq and its multi-modal variants (e.g. CITE-Seq, single-cell immune receptor profiling) have become the approaches of choice for an in-depth characterization of immune response phenotypes in the discovery of diagnostic, prognostic, and response biomarkers in vaccination, infection, and cancer studies (e.g. as reviewed in (Bode et al., 2021; Liu et al., 2021; Zielinski et al., 2021). SCHNAPPs is applicable for the analysis of data obtained by any of these approaches and can thus be integrated with most immunological studies today.

We exemplify the functionality of SCHNAPPs by characterizing human T cell phenotypes. We have performed CITE-Seq of peripheral blood mononuclear cells (PBMC). This bi-modal single-cell analysis enables simultaneous quantification of the full transcriptomics and a selected surface protein-expression profile for each cell (Stoeckius et al., 2017). The PBMC were extracted from the blood of two healthy individuals. The protein expression values (normalized and raw) for the PBMC experiment were added as projections in the data preparation step. In addition, we have performed scRNAseq on sorted CD4+ and CD8+ cells ("isolated T-cells" later in the text) from the same individuals to allow comparing the RNA-Seq data of PBMC with those of targeted T cells. The samples of the two individuals are labeled d1_ADT and d2_ADT for the PBMC and s1 and s2 for the isolated T-cells.

We have applied SCHNAPPs to identify major immune subsets in PBMC samples based on their transcriptional signatures. We have then correlated the RNA expression with the cell surface protein expression (ADT) to further characterize T cell subsets. We subsequently integrated T cells from PBMC with isolated T cells. This allowed us to infer the identity of T cell subsets in the isolated T-cells and to compare their gene expression profiles. Each step is accompanied by different visualizations.

In the following, we describe a concise workflow that shows the key features available in SCHNAPPs, with a step-by-step guide in the supplementary material ("paper-walkthrough"). Additional options and some alternative routes of investigation are given in the supplementary document "paper-walkthrough with alternatives". An in-depth walk-through and comparison to a scran workflow (https://bioconductor.org/packages/release/bioc/ vignettes/scran/inst/doc/scran.html) and a Seurat workflow (https://c3bi-pasteur-fr.github.io/UTechSCB-SCHNAPPs/articles/pkdown/SeuratWorkflow.html) are given in the supplementary material.

3.3. Preparing input for SCHNAPPs (0. paper-walkthrough)

As pointed out above, it consists of a SingleCellExperiment object in R with all available metadata saved in an ".RData" format. Alternatively, a simple count table either in a tab (.tsv) or comma (.csv) separated format can be used.

Any value that is independent of the ensemble of cells should be precomputed. This includes in our case reads aligning to mitochondrial sequences, gene annotations, protein expression from CITE-seq (raw and normalized), cell cycle prediction, and cell type predictions. An example of how to set up the data is given in the supplementary material (dataPrep.R) along with the resulting file (pbmc2020.RData).

3.4. Launching SCHNAPPs and loading input data (1. paper-walkthrough)

SCHNAPPs is started from RStudio's console using "schnapps(historyPath = '/Volumes/extDrive1/schnapps_history/PBMCExperiment')". The historyPath argument specifies the path to store major manipulations of the data and any plot that is requested to be stored. The date and time when the application is launched are recorded in the name of the subdirectory that holds all the files.

Using the input panel, the file to be analyzed is selected on the file system (pbmc2020.subsampled.RData). For this demonstration and to reduce the hardware requirements, we used SCHNAPPs to subsample the data to 1000 cells per sample. This is done by checking the "sub-sample" check-box as described in "paper-walkthrough". Thus, this example-analysis can be performed on a computer with 16GB of RAM.

3.5. Quality control (2. paper-walkthrough)

Several visualizations within SCHNAPPs are available for data quality control. The number of cells per sample (*General QC - Sample histogram*) allows identifying biases arising from different cell counts. Since we have subsampled the data this shows an even distribution of about 1000 cells per sample. In case of large differences of cell counts between the samples, one can sub-sample to a given maximum number of cells per sample (*input - input options - sub-sample*). When looking at the distribution of reads (*General QC - UMI histogram*, Fig. 1A) two distributions/modes can be seen that correspond to the two experimental approaches used (PBMC vs. isolated T-cells). This indicates potential biases in the data set that should be kept in mind when performing the analysis. One possibility of dealing with this bias would be to separate the four populations and analyze them independently. This visualization is also used to estimate the lower and upper boundaries for UMIs per cell [300,5000]. Cells with a very high number of UMIs can be considered as potential doublets and may be removed.

3.6. Identify cells with high mitochondrial content (3. paper-walkthrough)

Next, we remove cells with high mitochondrial content, which has been precalculated and is stored in the projection called "percent.mito". Cells with a high percentage of mitochondrial reads are most likely dying cells, where the cell membrane is already lyzed, leading to loss of the RNA and leaving mostly mitochondrial mRNA detectable in the cell (Ilicic et al., 2016; Islam et al., 2014). A value of 15% represents a relaxed threshold but corresponds visually with the distribution of the whole data set (Fig. 1B). This is done in the *co-expression - selected* tab by selecting the cells to be removed manually using the mouse and copying the cell identifiers to memory.

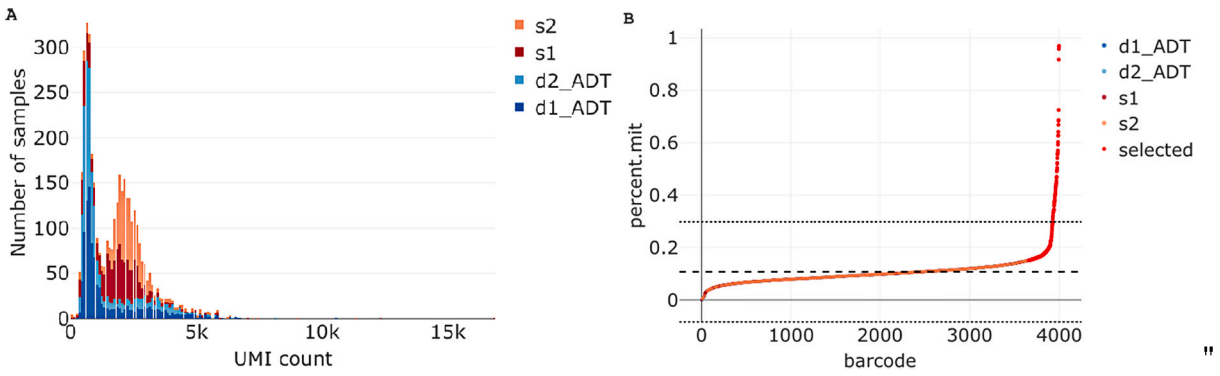

Fig. 1. Quality control figures of different steps in the quality control workflow. (A): Histogram of UMIs per sample (*General QC - UMI histogram*). PBMC (light (donor 2)/dark blue (donor 1)) and isolated T-cells (light (donor 2)/dark red (donor 1)). (B): 2D - plot (*Co-expression - selected*). With *barcode* on the x-axis, SCHNAPPs sorts the cells by the values on the y-axis. This allows for easier thresholding, especially when zooming is used (not shown). Cells with more than 15% mitochondrial content are selected (red dots) to be removed from the downstream analysis. Dotted lines show the 99.7 (3*σ) confidence interval, and the dashed line represents the mean. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.7. Remove cells and focus on PBMC (4. paper-walkthrough)

To remove these cells from the active data set they must be pasted under *Cell selection - cells to be removed*. Low and upper UMI thresholds are applied here as well. To remove the isolated T-cells and only keep the PBMC, the regular expression "s1|s2" is entered under *cells to be filtered out by pattern*. This takes advantage of the sample preparation step that added the sample name to each cell barcode.

We used this setup to show how one can start with a full data set, separate individual samples or groups of samples, analyze them separately, and then merge the results. This approach is often necessary to understand individual biases and to improve the biological interpretability of the data.

3.8. Normalization, dimension reduction, and clustering

Any single-cell analysis is very sensitive to the selection of variable features used for the PCA. Thus, different algorithms for calculating the PCA are implemented in SCHNAPPs to ensure that one can get the same results as with the Seurat or Scran pipelines. The intention of the next three steps is the following: normalization should render per-cell gene expression levels comparable; clustering should group cells in a biologically meaningful way; dimension reduction is used to visualize the expression and variance of the cells. These calculations must be activated on the *input* page by checking "*calculate logcounts using SCHNAPPs*". More information on the parameters for these steps (*Parameters - Cluster Parameters*) will be given later when data is integrated back with the scRNAseq data. For the analysis of PBMC in our example the standard options are sufficient.

3.9. Cluster characterization (5. paper-walkthrough)

Clustering and dimensionality reduction methods use different parameters that are intertwined. Hence, it is important to optimize the parameters together and verify that the results concur and are biologically meaningful. The 3D plot under *Parameters - tSNE plot* (Fig. 2A) allows for that. By changing the color from *sampleNames* to *dbCluster* it can be verified that the differences between samples are not dominating (*color = sampleNames, not shown*), i.e. the samples are not separated, but that clusters are visually separated in the dimensionality reduced space (*color = sampleNames, x,y,z = tSNE1/2/3*). Since clusters overlap and are often plotted on top of each other in two dimensions, the 3D view is important to better validate the calculations.

Different visual representations are available in SCHNAPPs to aid in

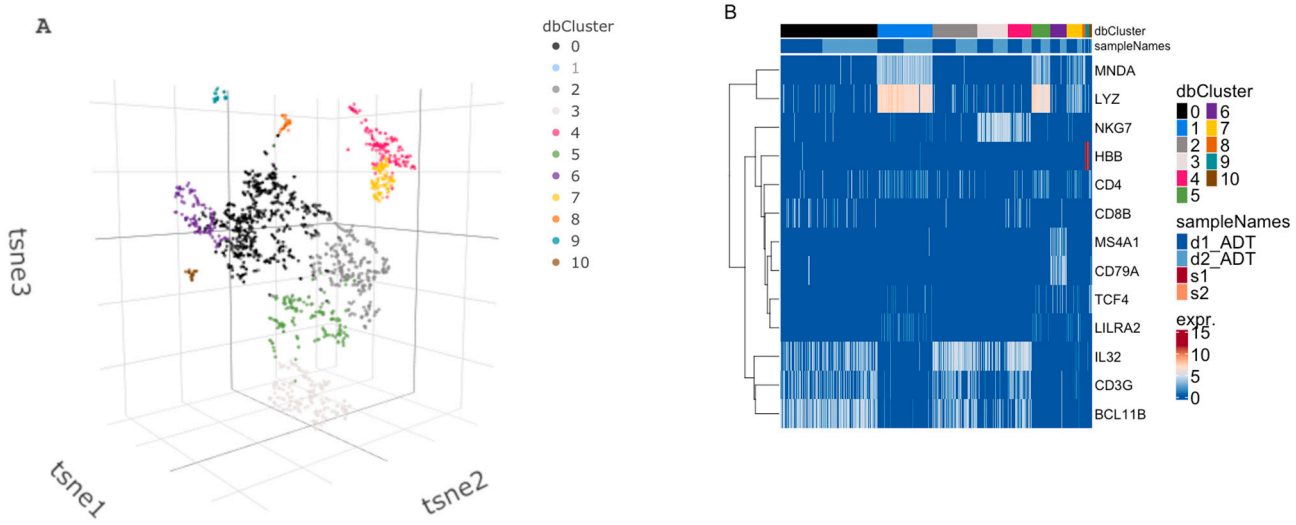

Fig. 2. Cluster validation. (A) 3D tSNE plot (*Parameters - TSNE plot*) showing PBMC cells in the 3D tSNE space colored by cluster identifier. All clusters are visually separated in the space. (B) Heatmap (*Co-expression - All clusters*) showing PBMC with marker genes that allow assigning biologically meaningful names (cell types) to clusters. The comparison between *dbCluster* and *sampleNames* in the top annotation shows that the samples are evenly distributed over the clusters.

characterizing clusters. Probably the most useful one is the heatmap (Fig. 2B). The comparison between *dbCluster* and *sampleNames* in the top annotation shows that the samples are evenly distributed over the clusters. This visualization also allows to visually compare the correlation of multiple markers of a given cell type with cluster assignments. Specifically, T cell marker *CD3G* correlates with clusters 0, 2, and 4; NK cells express no *CD3G* but *NKG7* as in cluster 3, while plasmacytoid dendritic cell (pDC) markers *TCF4*, *LILRA2* correlate with cluster 10, and cluster 6 expresses the B-cell gene markers *CD79A* and *MS4A1*. Clusters 1, 5, and 7 represent monocytes as they express *LYZ* and *MNDA*. Cluster 9 strongly expresses *HBB*, which indicates contamination by erythrocytes. Cluster 8 is difficult to characterize with the selection of genes.

3.10. Unbiased cluster characterization (6. paper-walkthrough)

Previously, prior knowledge about cell types was used to characterize the clusters. In cases where this is missing, SCHNAPPs calculates representative genes when the list of gene names is empty (using findMarkers from the scran package). This search can be restricted to genes that are differentially upregulated (*direction* = *up*) with a *minimum log fold change* of 2. The resulting list includes some of the genes used in Fig. 2B, highlighted in bold (*MALAT1*, *BCL11B*, *ETS1*, *BTG1*, *NSA2*, *LYZ*, *CTSS*, *S100A9*, *S100A8*, *VCAN*, *CD52*, *IL32*, *NEAT1*, *TOMM7*, *GNLY*, *NKG7*, *HLA-B*, *TMSB10*, *HLA-DRA*, *CD74*, *GZMA*, *IGHM*, *EEF1A1*, *PSAP*, *LST1*, *SERPINA1*, *COTL1*, *B2M*, *GATA2*, *TMSB4X*, *HBB*, *HBA2*, *HBA1*, *SLC25A37*, *NPC2*). These genes make biological sense and ensure the correctness of the cluster annotation. We can also infer that cluster 8 highly expresses *GATA2*, a gene expressed in granulocytes. All genes used for the initial heatmap can be found when loosening the parameters.

3.11. Apply knowledge to label clusters (7. paper-walkthrough)

Once the cell types represented by the clusters have been identified, the clusters can be renamed accordingly. This is done using the *Parameters - projections - rename levels* tab by creating a new projection (here, "manCellTypes").

3.12. Panel plot: project gene expression of selected genes on a 2D projection (8. a, b paper-walkthrough)

The panel plot (*Data exploration - panel plot*) visualizes the expression of individual marker genes in a 2D projection. This allows comparing genes in the context of a projection like tSNE, PCA, or UMAP. For example, it is possible to focus on B- and T-cells only (*Clusters/Factor to use* = *manCellTypes; Values to use* = *Bcell Tcell*) and plot the expression of six genes (*Comma separated gene names* = *CD3G*, *IGHM*, *CD74*, *BCL11B*, *IL32*, *NKG7*) on the first two tSNE axes (Fig. 3A). The normalized expression values are scaled between zero and one to better visualize lower expressed genes when highly expressed genes are present (same scale = unselected). The values are also sorted to ensure that the highest expressing cells are shown. This sorting can lead to an artefact when there are a few highly expressing cells among most cells that don't express this gene. In this case, most cells would not be visible. A more

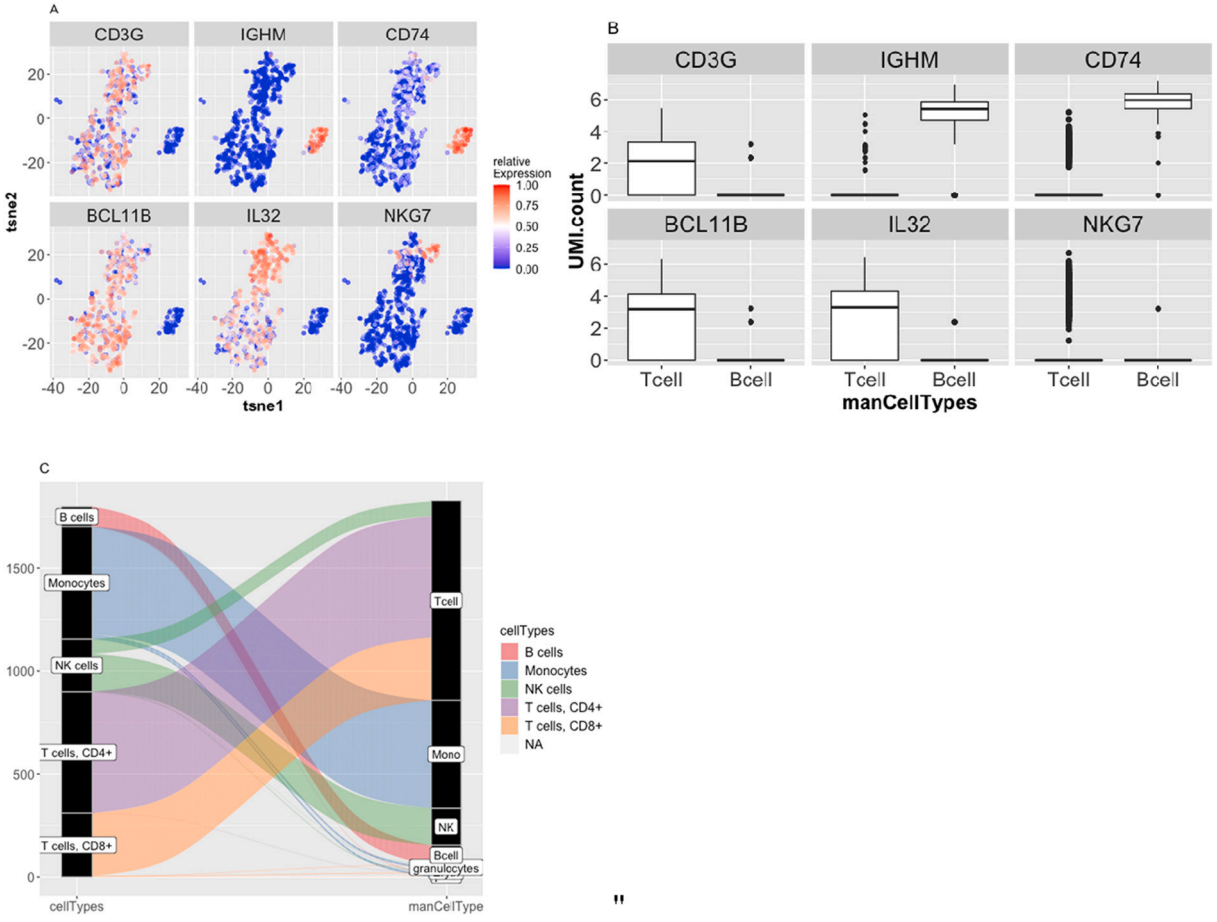

Fig. 3. (A): panel plot (*Data exploration - Panel plot*) of the relative normalized expression for selected genes. Only cells belonging to B- and T-cell clusters are projected on the tSNE coordinates. By un-selecting *same scale* in *Data exploration - Panel plot* the expression values are scaled per gene to a range between zero and one. (B): Using the same data as in A, the normalized expression is plotted in a box-and-whisker plot per gene. This allows for a more quantitative visualization. (C): Alluvial plot (*Co-expression - alluvialTab*) comparing singleR predicted (left) with manually annotated (right) cell types.

quantitative view that avoids this problem can be achieved by using a factorial variable (e.g., *manCellTypes*) on the X-axis and *UMI.count* on the Y axis, which results in a box-and-whisker plot (Fig. 3B). The non-scaled values are shown, allowing a direct comparison of mRNA expression, which is not controlled for mRNA length. These plots are used to validate cluster assignments and investigate how genes are expressed across different regions of the data set.

3.13. Alluvial plot (8.c paper-walkthrough)

The alluvial plot (*Co-expression - alluvialTab*) visualizes the correlation between manually identified cell types (manCellTypes) and cell types predicted by SingleR (Aran et al., 2019) during the setup of the data (Fig. 3C). Annotations correspond, except for a proportion of NK cells. Since effector T cells and NK cells share high expression of effector genes, they may be mis-assigned to the NK cell cluster.

3.14. Re-integrate isolated T-cells (9.-13. paper-walkthrough)

The following steps are detailed in the supplementary material (paper-walkthrough). In short, to analyze T-cells from all four experiments together only cells expressing *CD3* are used, also taking into account previous annotation. Isolated T-cells are re-integrated by removing the filtering on barcode names, Seurat's scTransform and anchor-based integration are applied, then visually validated and the new projections are copied by adding "sc" to the tSNE, UMAP, dbCluster variables, before restoring the standard log-transformed and UMI-count normalized expression values. This standard workflow follows Stuart et al. (Stuart et al., 2019) and allows visualizing the four data sets in common reduced dimensions. The most important parameters for each of these steps are tunable. It is not the purpose of this work to explain them. Optimizing the parameters is an iterative and project-specific process.

3.15. Working with integrated data (14. paper-walkthrough)

Basic analyses of cell populations can be performed using histograms, violin plots, or heatmaps. Here, we take advantage of the added value of protein expression data (CITE-Seq) to illustrate three different types of visualization. While gene expression provides information about *CD8* expression, and thus allows focusing the analysis on CD8+ T cells (see paper-walkthrough), protein expression is necessary to explore effector and naive subsets of T cells. Fig. 4A shows one way of visualizing the normalized protein expression of CD45RA and CD45RO under *Data Exploration - Expression*. It corresponds to the commonly used dot-plot representation of the flow cytometry protein expression data and allows for identification of effector vs. naive subsets of T cells. Here, only the manually assigned T-cells are plotted (Clusters/Factors to use = manCellTypes; Values to use = Tcell). This view was also used to define the cell groups named "ROpos", "RApos" that are stored in a new projection called RaRoCells using the previously described process for combining and renaming projections.

3.16. Counting cells (14.a paper-walkthrough)

Fig. 4B shows a barplot that visualizes the fraction of naive (CD45RA+) and effector (CD45RO+) subsets in both PBMC samples. Individual fraction values are accessible using the hover functionality of plotly-R (Sievert, 2020). Fig. 4C visualizes the RO+ vs. RA+ subsets that were defined based on protein markers in the tSNE space that is defined by mRNA expression. The distinction between RA+ and RO+ using the tSNE visualization is less clear compared to the protein data (Fig. 4A). Nonetheless, there is a visible enrichment on both ends of sctSNE-1 for RO+ (negative values) and RA+ (positive values). Because the sctSNE projections are calculated using all four data sets and there was no obvious sample specific bias in the projections, we can test if transferring the knowledge about RA/RO positive cells to the isolated T-cell experiment is biologically meaningful.

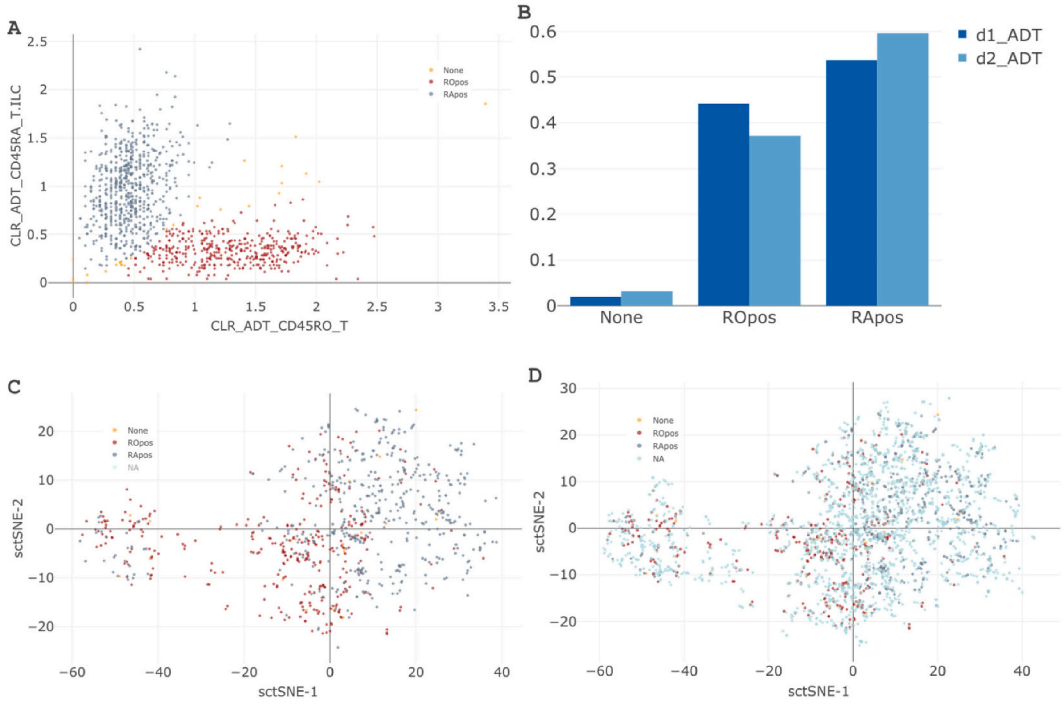

Fig. 4. (A): PBMC projected on the normalized protein expression of CD45RO and CD45RA. ROpos (red) and RApos (blue) have been manually assigned using the mouse. (B): Ratios of RO positive and RA positive cell counts are presented using a bar plot. Counts are normalized by the total number of cells per sample. (C): ROpos and RApos are highlighted in the integrated tSNE projection (sctSNE-1, sctSNE-2). The isolated T-cells have no RA/RO label and are marked as NA. They are unselected in the plot to show only the PBMC data. (D): Same as C, only showing all data, including isolated T cells.

Fig. 4D shows the combined data set of all four samples in sctSNE space, which was used to manually select the regions using the mouse and naming them ROpos, RApos, and NA (intermediate region). These new variables are combined into a new projection called ro.ra.all (see paper-walkthrough for further details).

3.17. Differential gene expression analysis (15. paper-walkthrough)

To study the difference between these selections the ROpos labeled cells are selected on the left-hand side of *Subcluster analysis - DGE analysis - Select data* (Fig. 5A). The RApos labeled cells are selected on the right side (Fig. 5A). The selection can be done with any 2D representation by manually gating on interesting groups of cells. This compares only the selected cells using a *t*-test (Method to use seurat:t-test). The results can be represented as a Volcano plot (Fig. 5B) or in a table with *p*-values, log-fold changes, and gene descriptions. Genes can be manually selected in the volcano plot and the gene names can be copy/pasted to any other visualization. CD45RO+ cluster contains memory and effectors T cells that downregulate CCR7 gene and upregulate effector-related genes such us the killer cell lectin like receptor G1 (KLRG1), cytotoxic molecules like Granzymes (GZMB, GZMH, and GZMA, GZMK), Perforin (PRF1), Granulysin (GNLY), NK cell granule protein (NKG7), chemokines like CCL5 and CCL4 and exhaustion markers like LAYN.

3.18. Co-expression (16. paper-walkthrough)

Using the violin plot panel under Co-expression one can visualize the co-occurrence of a given set of genes. The example represented in Fig. 5C shows the co-expression of any combination of *S100A4, CCL5, CCR7*, and *AIF1* for the ROpos and RApos labeled cells. This gives a unique way of visualizing the co-occurrence of multiple genes in relation to a factorial.

4. Discussion

In conclusion, we have applied the SCHNAPPs tool in characterization of immune cell subsets from human PBMC (CITE-Seq) and isolated T cells (scRNAseq) and showed how it allows a non-computer expert to gain valuable knowledge about single-cell RNA seq and multi-omics single cell experiments.

We have illustrated some of the main features that would be used first when analyzing scRNAseq data. The 2D plot alone, with its potential to show any combination of projections, meta-data, and groupings, allows for many quality-control and cell selection opportunities. In addition to the "contributions" discussed above, there are tools for investigating subsets of cells on what they have in common by calculating the coefficient of variance or correlation coefficients and weighted indexes for clustering evaluation (Wu and Wu, 2020). The heatmap also enables selecting cells, adjust the color pallets, create

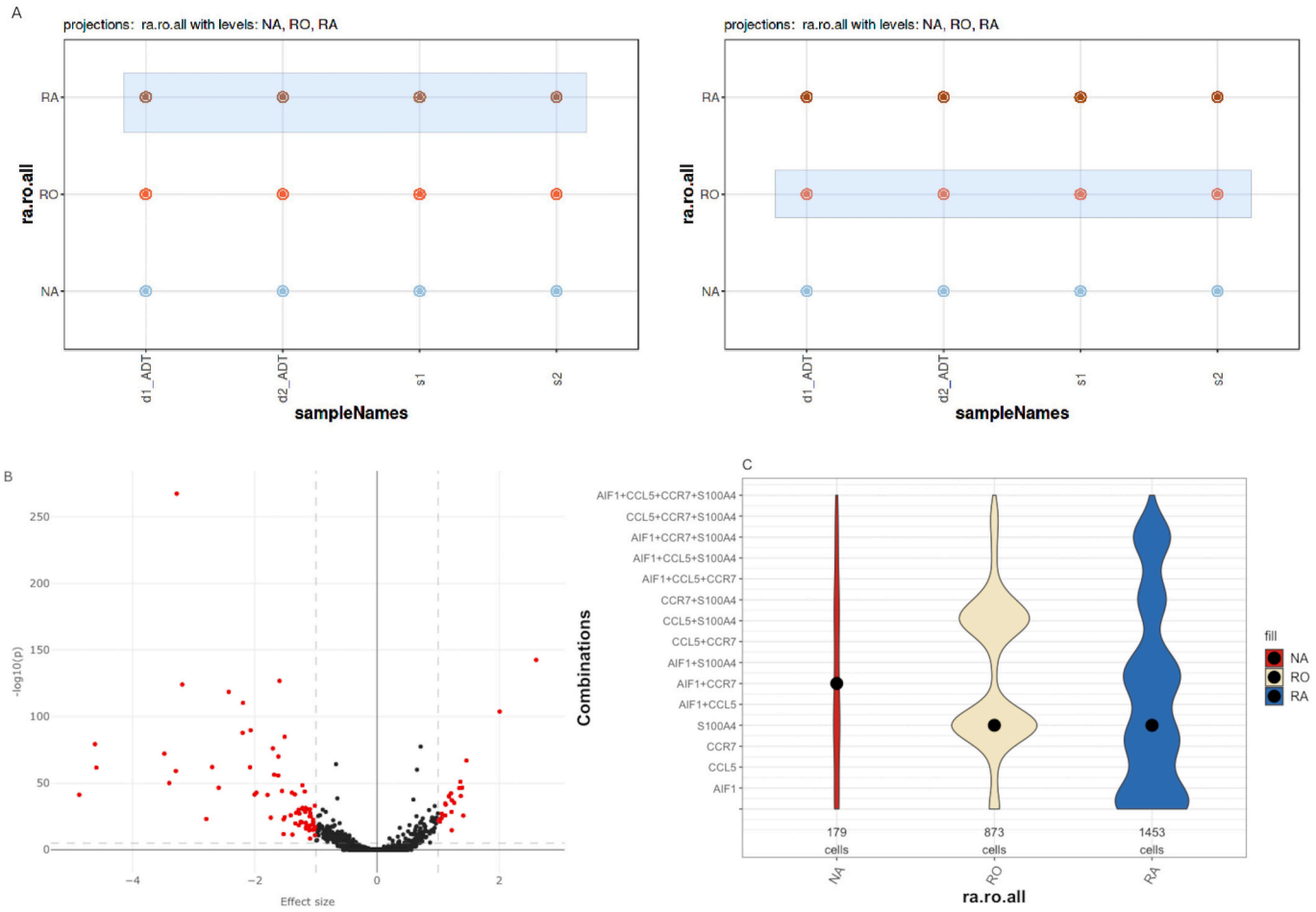

Fig. 5. (A): Cell selection for differential gene expression analysis (DGE, *Subcluster analysis - DGE analysis*). Cells are selected using the mouse. The comparison is done between the ROpos (left panel) and RApos subsets (right panel) of the integrated data set. (B): Volcano plot showing the result from the DGE in (A). Seurat's implementation of the t-test was selected when calculating the DGE. Genes can be selected using the mouse and are then displayed separately, which allows copying them to be used in other plots. (C): Violin plot (*Co-expression - Violin plot*) showing how many cells express a gene or combination of genes for the individual samples. Violins are scaled to have the same area.

dendrograms, use different scaling, and sort cells by expression of specific genes. Several advanced features of the SCHNAPPs, such as the option to define 'when a gene is being expressed' using thresholds for the lower and upper expression values in some visualizations, or the history functionality that ensures documenting a workflow and saving intermediate steps and visualizations, have not been detailed. To guide the user through basic and advanced functionalities, there are several vignettes, FAQs, and other information on GitHub (https://c3bi-pasteur-fr.github.io/UTechSCB-SCHNAPPs/articles/pkdown/SCHNAPPs_usage.html#co-expression).

SCHNAPPs has been successfully applied to validate a human "in a dish" model for a valvular disease (Neri et al., 2019) and to show that epicardium activation during a cardiomyopathy gives rise to both adipocytes and fibroblasts (Suffee et al., 2020). It is a standard tool for the analysis of scRNAseq data at the core facility Cytometry and Biomarkers UTechS at the Institut Pasteur.

The SCHNAPPs application is constantly evolving to integrate pipelines for other multi-omics data and its ongoing development is guided by individual use cases. In the near future, we plan to extend SCHNAPPs for the use of ATAC-seq (Buenrostro et al., 2015) data and optimize some of the visualizations.

We have also developed the function schnappsLite to enable publishing precomputed results. In this version of the tool, the compute-intensive components (normalization, dimension reduction, and clustering) have been removed and the number of cells can be limited to render publication results easily available to the general public. See http://hub05.hosting.pasteur.fr/scProjects/ for examples.

In summary, SCHNAPPs provides a framework for reproducible, methodologically sound exploration of scRNAseq datasets with interactive visualizations.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jim.2021.113176.

Funding

This work has been supported by the Institut Pasteur.

Availability

The SCHNAPPs application is available as an R-package (https://github.com/C3BI-pasteur-fr/UTechSCB-SCHNAPPs), docker file (https://hub.docker.com/r/pf2dock/schnapps), and virtual machine (doi:https://doi.org/10.5281/zenodo.5535294). Extensive documentation including videos, walkthroughs, and frequently asked questions are available on GitHub (https://c3bi-pasteur-fr.github.io/UTechSCB-SCHNAPPs/) and youtube: https://tinyurl.com/schnappsYT. Example contributions are available at the following GitHub site: github.com/C3BI-pasteur-fr/SCHNAPPsContributions. All raw sequencing data are available from the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE185712).

Bode, D., Cull, A.H., Rubio-Lara, J.A., Kent, D.G., 2021. Exploiting Single-Cell Tools in Gene and Cell Therapy. Front Immunol 12, 702636. doi:https://doi.org/10.3389/fimmu.2021.702636

Declaration of Competing Interest

SCHNAPPs has been patented under: IDDN1.FR2 0.0013 0.3600164 0.0005.S6.P7 0.20208 0.0009 0.3123510.

Acknowledgments

We would like to thank the members of Single-cell working group Pasteur/Paris for helpful discussions: Anna Barcons, Eric Tartour, Antonin Saldmann, Mandar Patgaonkar, Lisa Chakrabarti, and James Di Santo for testing and working with scShinyHub and SCHNAPPs. Kenneth Smith and Christian Vosshenrich for careful reading of the manuscript. We thank the ICAReB platform of the Institut Pasteur for providing blood samples from healthy individuals.

References

Albergante, L., 2021. ElPiGraph.R: Elastic Principal Graph Construction.
Amezquita, R.A., Carey, V.J., Carpp, L.N., Geistlinger, L., Lun, A.T.L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M., Huber, W., Morgan, M., Gottardo, R., Hicks, S.C., 2019. Orchestrating Single-Cell Analysis with Bioconductor. https://doi.org/10.1101/590562. http://web.archive.org/web/20200503103403/https://www.biorxiv.org/content/10.1101/590562v1.
Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., Bhattacharya, M., 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. 20, 163–172. https://doi.org/10.1038/s41590-018-0276-y.
Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr. Prot. Mol. Biol. 109 https://doi.org/10.1002/0471142727.mb2129s109, 21.29.1-21.29.9.
Cakir, B., Prete, M., Huang, N., van Dongen, S., Pir, P., Kiselev, V.Y., 2020. Comparison of visualization tools for single-cell RNAseq data. NAR Gen. Bioinformat. 2 https://doi.org/10.1093/nargab/lqaa052.
Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D., Saeys, Y., 2016. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development (preprint). Bioinformatics. https://doi.org/10.1101/079509.
Chang, W., Borges Ribeiro, B., 2018. shinydashboard: Create Dashboards with "Shiny".
Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2021. shiny: Web Application Framework for R.
Cortal, A., Martignetti, L., Six, E., Rausell, A., 2021. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nat. Biotechnol. 39, 1095–1102. https://doi.org/10.1038/s41587-021-00896-6.
Donaldson, J., 2016. tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE).
Duan, B., Zhu, C., Chuai, G., Tang, C., Chen, X., Chen, S., Fu, S., Li, G., Liu, Q., 2020. Learning for single-cell assignment. Sci. Adv. 6, eabd0855. https://doi.org/10.1126/sciadv.abd0855.
Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., Theis, F.J., 2019. Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. 10, 390. https://doi.org/10.1038/s41467-018-07931-2.
Gardeux, V., David, F.P.A., Shajkofci, A., Schwalie, P.C., Deplancke, B., 2017. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 33, 3123–3125. https://doi.org/10.1093/bioinformatics/btx337.
Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
Hillje, R., Pelicci, P.G., Luzi, L., 2020. Cerebro: interactive visualization of scRNA-seq data. Bioinformatics 36, 2311–2313. https://doi.org/10.1093/bioinformatics/btz877.
Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., Teichmann, S.A., 2016. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 17, 29. https://doi.org/10.1186/s13059-016-0888-1.
Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11, 163–166. https://doi.org/10.1038/nmeth.2772.
Jenkins, D.F., Faits, T., Briars, E., Pro, S.C., Cunningham, S., Campbell, J.D., Yajima, M., Johnson, W.E., 2018. Interactive single cell RNA-Seq analysis with the Single Cell Toolkit (SCTK). https://doi.org/10.1101/329755.
Liu, J., Qu, S., Zhang, T., Gao, Y., Shi, H., Song, K., Chen, W., Yin, W., 2021. Applications of single-cell omics in tumor immunology. Front. Immunol. 12, 697412 https://doi.org/10.3389/fimmu.2021.697412.
McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [cs, stat].
Neri, T., Hiriart, E., van Vliet, P.P., Faure, E., Norris, R.A., Farhat, B., Jagla, B., Lefrancois, J., Sugi, Y., Moore-Morris, T., Zaffran, S., Faustino, R.S., Zambon, A.C., Desvignes, J.-P., Salgado, D., Levine, R.A., de la Pompa, J.L., Terzic, A., Evans, S.M., Markwald, R., Pucéat, M., 2019. Human pre-valvular endocardial cells derived from pluripotent stem cells recapitulate cardiac pathophysiological valvulogenesis. Nat. Commun. 10, 1929. https://doi.org/10.1038/s41467-019-09459-5.
Patel, M.V., 2018. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics 34, 4305–4306. https://doi.org/10.1093/bioinformatics/bty517.
RStudio Team, 2020. RStudio: Integrated Development Environment for R. RStudio. PBC, Boston, MA.
Rue-Albrecht, K., Marini, F., Soneson, C., Lun, A.T.L., 2018. iSEE: interactive summarizedexperiment explorer. F1000Res 7, 741. https://doi.org/10.12688/f1000research.14966.1.
Sievert, C., 2020. Interactive Web-Based Data Visualization with R, Plotly, and Shiny. CRC Press.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P., 2017. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14, 865–868. https://doi.org/10.1038/nmeth.4380.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R., 2019. Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Suffee, N., Moore-Morris, T., Jagla, B., Mougenot, N., Dilanian, G., Berthet, M., Proukhnitzky, J., Le Prince, P., Tregouet, D.A., Pucéat, M., Hatem, S.N., 2020. Reactivation of the epicardium at the origin of myocardial fibro-fatty infiltration during the atrial cardiomyopathy. Circ. Res. 126, 1330–1342. https://doi.org/10.1161/CIRCRESAHA.119.316251.

Tran, T.N., Bader, G.D., 2019. Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. https://doi.org/10.1101/846907. http://web.archive.org/web/20200426152125/https://www.biorxiv.org/content/10.1101/846907v1.

Wu, Z., Wu, H., 2020. Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. Genome Biol. 21, 123. https://doi.org/10.1186/s13059-020-02027-x.

Xie, Y., Allaire, J.J., Grolemund, G., 2019. R Markdown: the Definitive Guide. CRC Press, Taylor and Francis Group, Boca Raton.

Yates, A.D., Achuthan, P., Akanni, W., Allen, James, Allen, Jamie, Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J.C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S. H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J.G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D.N., Parker, A., Parton, A., Patricio, M., Sakthivel, M.P., Abdul Salam, A.I., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S.E., IIsley, G., Kostadima, M., Langridge, N., Loveland, J.E., Martin, F.J., Morales, J., Mudge, J.M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S.J., Cunningham, F., Howe, K.L., Zerbino, D.R., Flicek, P., 2020. Ensembl 2020. Nucleic Acids Res. 48, D682–D688. https://doi.org/10.1093/nar/gkz966.

Yousif, A., Drou, N., Rowe, J., Khalfan, M., Gunsalus, K.C., 2020. NASQAR: a web-based platform for high-throughput sequencing data analysis and visualization. BMC Bioinformat. 21, 267. https://doi.org/10.1186/s12859-020-03577-4.

Zielinski, J.M., Luke, J.J., Guglietta, S., Krieg, C., 2021. High throughput multi-omics approaches for clinical trial evaluation and drug discovery. Front. Immunol. 12, 590742 https://doi.org/10.3389/fimmu.2021.590742.