scOrange: Single-Cell Data Mining for Everyone

Martin Stražar,[1] Lan Žagar,[1] Jaka Kokošar,[1] Vesna Tanko,[1] Pavlin Poličar,[1] Aleš Erjavec,[1] Anže Starič,[1] Vilas Menon,[2] Rui Chen,[3] Gad Shaulsky,[3] Andrew Lemire,[2] Anup Parikh,[4] Blaž Zupan[1,3]

[1] University of Ljubljana, Ljubljana, Slovenia, [2] Howard Hughes Medical Institute, Ashburn, U.S.A.
[3] Baylor College of Medicine, Houston, U.S.A., [4] Naringi, San Francisco, U.S.A.

Introduction

Single-cell RNA sequencing (scRNA-seq) can profile gene expression at a single-cell resolution [1]. In comparison to bulk RNA-seq methods, it provides a detailed insight into cell populations from an organism or a tissue and enables tracking of disease onset to previously unknown cell subpopulations or genotypes. scRNA data can include thousands of cells and provides opportunities for visualization and modeling. The contemporary software packages for single-cell data analytics assume knowledge of programming and limit the analyses to computer-savvy individuals [2]. We present scOrange, a single single-cell extension of the data mining toolbox Orange that features interactive visualizations and dynamic data analysis workflows [3]. Our tool enables rapid prototyping of reusable data analysis workflows, promoting reproducibility and the use of standardized, literature-supported approaches. It assumes no programming knowledge and is accessible to a broader audience than present scripting libraries in R or Python.

Methods

In scOrange, the user assembles workflows by selecting and connecting data sources, data analysis, and visualization components (Fig. 1). Components in scOrange are interactive; any change in a component – say, by loading a different data set, change of a modeling parameter, or selection of a subset of genes or cells from a visualization – propagates to the downstream components. Through a specific combination of components in a workflow, scOrange users can address a particular problem and construct a customized exploratory data analysis environment. Workflows can be saved and shared, promoting transparency and reproducibility in analysis.

scOrange can load gene expression data from a variety of input file types (csv, xlsx, loom, mtx) and supports a range of data preprocessing techniques including scaling, normalization, filtering and scoring of cells and genes. The gene IDs are matched to standardized names in the NCBI database. Merging datasets from multiple protocols or experimental conditions consists of selecting shared genes and removing of batch effects by linear regression or canonical correlation analysis. Data preprocessing results may be assessed in interactive visualizations that use dimensionality reduction by PCA, MDS, or t-SNE. Visualisations can highlight clusters of subpopulations or cells that express a selected set of marker genes, and implement brushing and linking to support exploratory data analysis. Clustering methods include k-means, hierarchical clustering, and network-based approaches. Cell clusters may be explored through differential expression analysis and gene set enrichment in the space of labels from ontologies such as Gene Ontology or KEGG pathways.

Results

The scOrange tool is available in open source (http://singlecell.biolab.si). We have tested it by replicating recently published studies on different organisms and cell types. These include dataset alignment and batch effect removal [4], preprocessing and normalization, clustering and cluster-based differential expression [5], cell-cycle stage prediction [6], the discovery of new putative cell populations and their corresponding markers [7]. Studies that include up to tens of

thousands of cells can be processed on a standard laptop and typically take only a few minutes of runtime. Accompanying resources include pre-defined workflows, and blog posts with case studies. Video tutorials are in development and are available on scOrange's homepage.

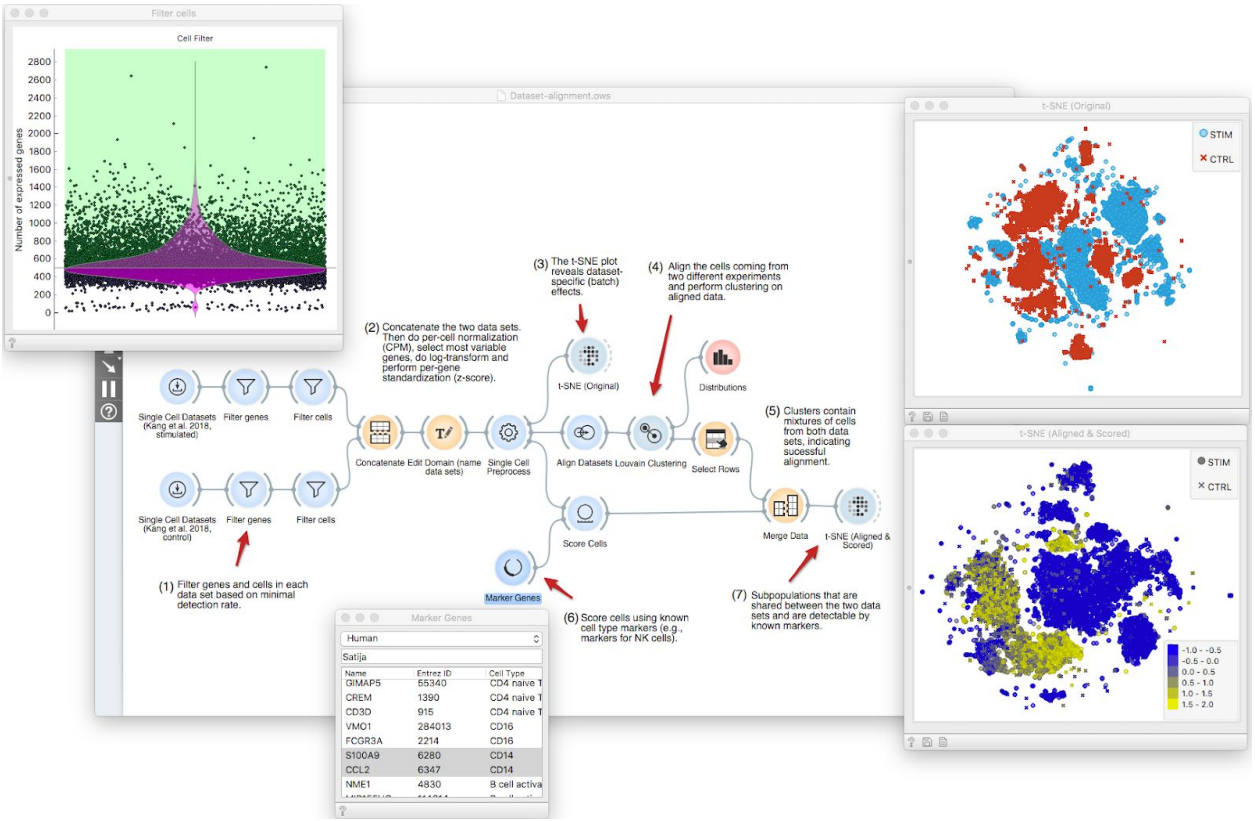

Fig. 1: An example workflow based on the study of Butler et al., 2018 [4]. The workflow includes, in order from left: loading data from two data sets; filtering and preprocessing; data set alignment (CCA); Louvain clustering and t-SNE visualization, augmented with the scoring of cells based on putative marker genes.

Conclusions

scOrange is a data mining toolbox that aims to democratize single cell data science with an intuitive user interface, interactive visualizations, and Lego bricks-like approach for workflow construction. By constraining the functionality to literature-supported methods, it alleviates the risk of overfitting to particular data sets or protocol. With the proliferation of scRNA-seq data, we envision that interactive visualizations and workflow construction through visual programming will become an integral part of biomedical research and promote FAIR principles in data analysis.

References

1. DA Jaitin, *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 6172: 776-779 (2014).
2. Alexander WF, Angerer P, and Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**.1, 15 (2018).
3. Demšar J, *et al.* Orange: data mining toolbox in Python. *The Journal of Machine Learning Research* **14**.1, 2349-2353 (2013).
4. Andrew B, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**.5, 411 (2018).
5. Alexandra-Chloé V, *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**.6335 (2017).
6. Antonio S, *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54-61 (2015).
7. Evan ZM, *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**.5: 1202-1214 (2015).