



Gene expression

tagtango: an application to compare single-cell annotations

Bernat Bramon Mora ^{1,2,3}, Helen Lindsay^{1,2,3}, Antonin Thiébaut^{1,2,3}, Kenneth D. Stuart ⁴,
Raphael Gottardo^{1,2,3,5,*}

¹Biomedical Data Science Center, Lausanne University Hospital, Vaud 1005, Switzerland

²Biomedical Data Science Center, University of Lausanne, Vaud 1015, Switzerland

³Swiss Institute of Bioinformatics, Vaud 1015, Switzerland

⁴Center for Global Infectious Disease Research, Seattle Children's Hospital, WA 98105, United States

⁵School of Life Sciences, EPFL - Swiss Federal Technology Institute of Lausanne, Lausanne, Vaud 1015, Switzerland

*Corresponding author. Biomedical Data Science Center (BDSC); Lausanne University Hospital; Bureau BU21 / 05 / 234 – Mail box 50; Rue du Bugnon 21, CH-1011 Lausanne. E-mail: raphael.gottardo@chuv.ch

Associate Editor: Macha Nikolski

Abstract

Summary: In this article, we present tagtango, an innovative R package and web application designed for robust and intuitive comparison of single-cell clusters and annotations. It offers an interactive platform that simplifies the exploration of differences and similarities among different clustering and annotation methods. Leveraging single-cell data analysis and different visualizations, it allows researchers to dissect the underlying biological differences across groups. tagtango is a user-friendly application that is portable and works seamlessly across multiple operating systems.

Availability and implementation: tagtango is freely available at <https://github.com/bernibra/tagtango> as an R package as well as an online web service at <https://tagtango.unil.ch>.

1 Introduction

Two integral components of single-cell data analysis are cell clustering and annotation (Kiselev *et al.* 2019). They are central to most downstream analyses (e.g. differential expression analyses; Finak *et al.* 2015), allowing us to study relevant biological processes such as transcriptional changes and cell/gene interactions (Kumar *et al.* 2018, van Dijk *et al.* 2018). This has led to the development of a diverse collection of clustering and annotation methods to automate the identification of cell populations (de Kanter *et al.* 2019, Alquicira-Hernandez *et al.* 2019, Hao *et al.* 2021). This diversity of methods and granularity of annotations, however, in combination with the heterogeneous nature of biological data, can lead to large variability in the identification of phenotypes (Ziegenhain *et al.* 2017). The emergence of multimodal single-cell sequencing technologies—e.g. surface protein expression (Stoeckius *et al.* 2017), DNA methylation (Gaiti *et al.* 2019), chromatin accessibility (Cao *et al.* 2018), and spatial transcriptomics (Ståhl *et al.* 2016)—has opened the door to applying such annotation methods to different data modalities, adding even more complexity to the identification and classification of cell types. Indeed, new data modalities could provide additional resolution to identify cell populations that cannot be distinguished solely based on RNA expression (Ding *et al.* 2020). Therefore, we need tools that help us compare annotations and clusters, untangling differences and similarities across populations. Here, we introduce

tagtango, an R package and user-friendly web application designed for comparing single-cell annotations and clusters. This tool offers an easy way to shed light on inconsistencies present across various cell identification methods and better understand whether the variations across annotations contain relevant biological information or are the product of the idiosyncrasies of different data types and methods.

2 Materials and methods

tagtango is a software package to compare multiple annotations associated to a single-cell dataset. The input data can be provided in standard Bioconductor formats (Gentleman *et al.* 2004): a ‘MultiAssayExperiment’ object stored as an R Data Serialization (RDS) file (Ramos *et al.* 2017); a ‘SingleCellExperiment’ object stored as an RDS file (Amezquita *et al.* 2020); or a data frame as an RDS, comma-separated values, or tab-separated values file. The core functionality of tagtango centres around running a web application that displays a detailed comparison of annotations generated by different methods and data types. That is, given an input data object containing a normalized expression matrix and a set of annotations, tagtango will study the differences in marker expression across cell populations (see [Supplementary Material](#) for data specifications). This application also allows the user to hone in on relevant differences across annotation strategies by strategically grouping

Received: 16 July 2024; Revised: 11 December 2024; Editorial Decision: 2 January 2025; Accepted: 8 January 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

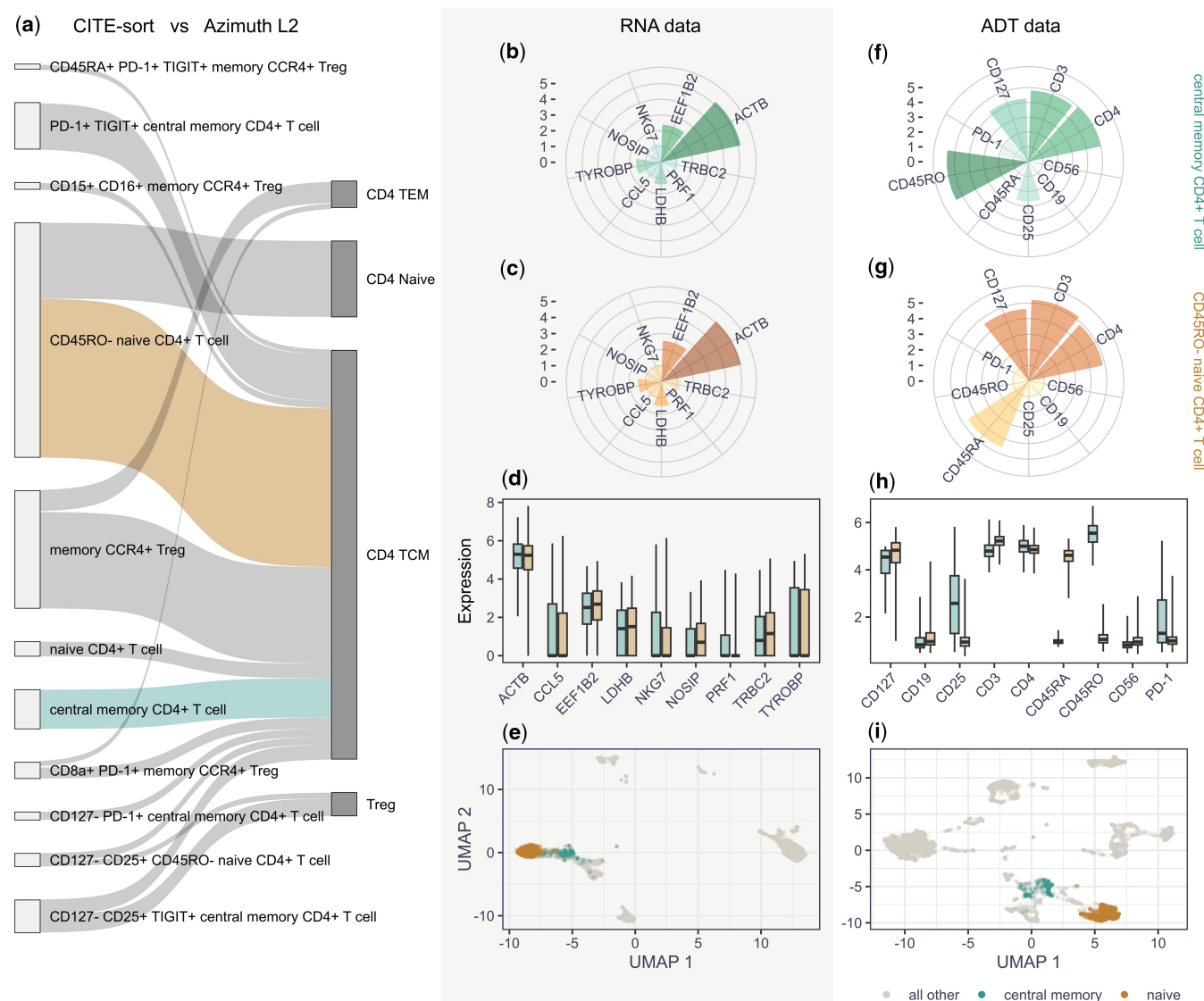


Figure 1. Overview of the annotation comparison performed by tagtango. Panel (a) displays a Sankey diagram comparing the annotations performed by CITE-sort and Azimuth's level 2 (i.e. 'celltype.l2'). The diagram was filtered using tagtango to only include cells annotated as 'CD4+ T-cells' by Azimuth's main cell type classification (i.e. 'celltype.l1') and links containing at least 20 cells. The coloured links in the diagram indicate the cell populations selected for deeper analysis. Panels (b) and (f) display rose plots of the normalized expression for the genes and protein markers deemed most relevant in the selected cell population 'central memory CD4+ T cell'. Panels (c) and (g) display the same for the selected cell population 'CD45RO- naive CD4+ T cell'. Panels (d) and (h) display a direct comparison between the ADT and RNA marker normalized expression for the two selected cell populations, including only those relevant markers. The colours of the bars match those of the selected links in panel (a). Panels (e) and (i) present the UMAP representation of all cells calculated using all markers for the ADT and RNA expression data, respectively. The colours of the points match those of the selected links in panel (a).

populations across main cell types or other variables, as well as filtering out noise or inconsistencies across them. Notably, tagtango allows users to export all results and figures as R code (see Fig. 1), ensuring reproducibility for every visualization produced by the tool.

Given two cell populations of interest, tagtango uses different strategies to select the relevant markers displayed in the visualizations. When working with low-dimensional data modalities, such as antibody-derived tags (ADT) expression data, tagtango identifies relevant protein markers by studying differences in average expression between populations, allowing the users to also specify other required markers in the visualizations. In contrast, when working with RNA expression data (or other high-dimensional data modalities; i.e. >1000 features), tagtango first uses the function

'scoreMarkers' from the R package 'scran' (Lun *et al.* 2016) to quantify the expression differences across every possible annotation or cluster, computing summary scores for each marker in each group of cells. Then, it selects the 10 most upregulated markers for every group using the median Cohen's *d*. Finally, it identifies the markers displayed in the visualization by studying the average expression differences between any selected populations, allowing the users again to include any other upregulated marker across possible annotations or clusters. Notice that, alternatively, one can provide filtered RNA expression data including only gene markers of interest, thus bypassing the need for the pre-selection using 'scran' and speeding up the analyses.

The implementation of tagtango involves an R package and a shiny application, which use Javascript and R libraries

such as ‘networkD3’ (Allaire *et al.* 2017) and ‘ggplot2’ (Wickham 2016) to generate the visualizations. The software, distributed under the MIT license, is cross-platform and freely available at <https://github.com/bernibra/tagtango>. The accompanying web application is hosted at <https://tagtango.unil.ch>, though, for cost reasons, it is currently restricted to smaller tests (i.e. approximately >10 000 cells and <2000 features). In terms of resource requirements, the test dataset (7472 cells) requires approximately 840 MB of memory to perform operations in R. Running tagtango on this dataset using the low-dimensional data approach (fewer than 2000 features) requires 1 CPU and approximately 1010 MB of RAM. For higher-dimensional data modalities, the internal use of ‘scoreMarkers’ by tagtango increases the RAM consumption up to 4.2GB for the same dataset (using 33 538 features). Both resource and time scalability will depend on the number of unique annotations or clusters and the use of ‘scoreMarkers’, which has been benchmarked in the past using multiple datasets (Pullin and McCarthy 2024).

3 Usage scenario: ADT vs RNA-based annotations

To illustrate the functionality of tagtango, we used the pre-processed and annotated dataset provided with the R package and web application. This is a ‘MultiAssayExperiment’ with peripheral blood mononuclear cells from a healthy donor, stained with TotalSeq-B antibodies (10x Genomics 2018). To normalize the RNA expression matrix, we used the ‘logNormCounts’ function from the R package ‘scuttle’ (McCarthy *et al.* 2017), computing therefore the log-transformed normalized expression values. Similarly, we normalized the ADT expression data using the R package ‘ADTnorm’ (Zheng *et al.* 2022), a tool specifically designed for the normalization of CITE-seq data. The dataset was annotated using different algorithms. In particular, we annotated cells using RNA expression data with Azimuth (Hao *et al.* 2021), and the ADT with CITE-sort (Lian *et al.* 2020).

Using tagtango, we compared the two sets of annotations. First, we grouped cells based on their main cell-type classification (Azimuth level 1) and selected CD4 T-cells as our main focus. With this information, tagtango produces a Sankey diagram, where nodes on each side represent the different annotations and links characterize their associations. Figure 1a shows the resulting diagram when filtering out links and nodes containing <20 cells (note that this parameter can be adjusted in the application).

The Sankey diagram allows the user to select any pair of nodes or links and compare the corresponding cell populations. In Fig. 1a, we selected two populations: cells annotated as ‘central memory CD4+ T cell’ by CITE-sort, and cells annotated as ‘CD45RO- naive CD4+ T cell’ by CITE-sort. Notice that these two populations appear somewhat inconsistent with each other, as the CITE-sort annotations differentiated two cell types while Azimuth classified them both as central memory CD4+ T-cells (i.e. ‘CD4 TCM’). Using the normalized ADT and RNA expression data, tagtango produced several visualizations to illustrate their differences. While the RNA visualizations found similar marker expression levels between annotations (Fig. 1b–e), the ADT visualizations identified markers CD45RA and CD45RO as the primary responsible for such inconsistencies (Fig. 1f–i). This seems to support the distinction made by CITE-sort, as Naive

CD4+ T cells should be characterized by the expression of CD45RA, which indicate cells that have not encountered antigens. Likewise, memory CD4+ T cells should exhibit CD45RO expression, indicating a history of antigen exposure and differentiation (Henson *et al.* 2012). Overall, this highlights the potential challenge that annotations relying on RNA expression may encounter in distinguishing between various isoforms of a given protein; in this case, two isoforms CD45RA and CD45RO resulting from the alternative splicing of the PTPRC gene. In contrast, however, both the RNA and ADT UMAP representations (Fig. 1e and i) seem to separate well the two cell populations, showing the cumulative effect of differences in the expression across genes.

4 Conclusion

As the landscape of single-cell technologies and annotation methods continues to grow, we need tools to compare and interpret annotations and clusters identified by different algorithms and across data modalities (Freytag *et al.* 2018, Xu *et al.* 2022). tagtango is an R package and web application developed to facilitate such comparisons. Taking a single-cell dataset as input, tagtango facilitates the analysis and synthesis of the differences and similarities across any set of annotations by producing visualizations comparing specific cell populations. In this work, we showcased its utility by comparing annotations produced using RNA and ADT expression data in a CITE-seq dataset. In particular, we were able to key in on specific inconsistencies across methods, shedding light on the potential pitfalls of each annotation strategy. Moreover, we provided additional examples of how tagtango can be used across data modalities (see the ‘Supplementary usage scenario: comparing spatial transcriptomics annotations’ section of the Supplementary Information) and for any type of annotation (including samples information as shown in the ‘Supplementary usage scenario: understanding batch effects’ section of the Supplementary Information). Likewise, we provided an example of cross-dataset comparisons using independent single-cell experiments with tagtango (see the ‘Supplementary usage scenario: comparing single-cell datasets’ section of the Supplementary Information). Overall, we believe that this new tool could be of great use for the single-cell community, enabling the comparison of annotations to become a routine part of an analysis, even for non-expert users.

Author contributions

Bernat Bramon Mora led the design of the work, developed the software, led the writing, and gave final approval for publication. Helen Lindsay, Antonin Thiébaud, and Raphael Gottardo contributed to the design of the work, contributed to the preparation of the manuscript, and gave final approval for publication. Kenneth D. Stuart contributed to the preparation of the manuscript and gave final approval for publication.

Supplementary data

Supplementary data are available in the R package at <https://github.com/bernibra/tagtango> and in the article’s online supplementary material.

Conflict of interest: R.G. has received consulting income from GSK, Takeda, Sanofi and Arcellx, and discloses

ownership in Ozette Technologies. Additionally, R.G. declares research collaborations with Owkin and 10X Genomics.

Funding

This work was supported by Chan Zuckerberg Initiative award DI-0000000345 and NIH award U19AI128914.

References

- Allaire J, Gandrud C, Russell K *et al.* *networkD3: D3 JavaScript Network Graphs from R*. 2017. R package version 0.4. <https://github.com/christophergandrud/d3Network>.
- Alquicira-Hernandez J, Sathe A, Ji HP *et al.* scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;20:264. <https://doi.org/10.1186/s13059-019-1862-5>
- Amezquita RA, Lun ATL, Becht E *et al.* Orchestrating single-cell analysis with bioconductor. *Nat Methods* 2020;17:137–45. <https://doi.org/10.1038/s41592-019-0654-x>
- Cao J, Cusanovich DA, Ramani V *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (New York, N.Y.)* 2018;361:1380–5. <https://doi.org/10.1126/science.aau0730>
- de Kanter JK, Lijnzaad P, Candelli T *et al.* CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;47:e95. <https://doi.org/10.1093/nar/gkz543>
- Ding J, Adiconis X, Simmons SK *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* 2020;38:737–46. <https://doi.org/10.1038/s41587-020-0465-8>
- Finak G, McDavid A, Yajima M *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;16:278. <https://doi.org/10.1186/s13059-015-0844-5>
- Freytag S, Tian L, Lönnstedt I *et al.* Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data. *F1000Res* 2018;7:1297. <https://doi.org/10.12688/f1000research.15809.2>
- Gaiti F, Chaligne R, Gu H *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* 2019;569:576–80. <https://doi.org/10.1038/s41586-019-1198-z>
- 10x Genomics. 10k PBMCs from a Healthy Donor, Single Cell Gene Expression Dataset by Cell Ranger 3.0.0. 2018.
- Gentleman RC, Carey VJ, Bates DM *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5:R80. <https://doi.org/10.1186/gb-2004-5-10-r80>
- Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multi-modal single-cell data. *Cell* 2021;184:3573–87.e29. <https://doi.org/10.1016/j.cell.2021.04.048>
- Henson SM, Riddell NE, Akbar AN. Properties of end-stage human T cells defined by CD45RA re-expression. *Curr Opin Immunol* 2012;24:476–81. <https://doi.org/10.1016/j.coi.2012.04.001>
- Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;20:273–82. <https://doi.org/10.1038/s41576-018-0088-9>
- Kumar MP, Du J, Lagoudas G *et al.* Analysis of single-cell RNA-seq identifies cell-cell communication associated with tumor characteristics. *Cell Rep* 2018;25:1458–68.e4. <https://doi.org/10.1016/j.cellrep.2018.10.047>
- Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;17:75. <https://doi.org/10.1186/s13059-016-0947-7>
- Lian Q, Xin H, Ma J *et al.* Artificial-cell-type aware cell-type classification in CITE-seq. *Bioinformatics* 2020;36:i542–i550. <https://doi.org/10.1093/bioinformatics/btaa467>
- McCarthy DJ, Campbell KR, Lun ATL *et al.* Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 2017;33:1179–86. <https://doi.org/10.1093/bioinformatics/btw777>
- Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol* 2024;25:56. <https://doi.org/10.1186/s13059-024-03183-0>
- Ramos M, Schiffer L, Re A *et al.* Software for the integration of multiomics experiments in bioconductor. *Cancer Res* 2017;77:e39–42. <https://doi.org/10.1158/0008-5472.CAN-17-0344>
- Ståhl PL, Salmén F, Vickovic S *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)* 2016;353:78–82. <https://doi.org/10.1126/science.aaf2403>
- Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;14:865–8. <https://doi.org/10.1038/nmeth.4380>
- van Dijk D, Sharma R, Nainys J *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;174:716–29.e27. <https://doi.org/10.1016/j.cell.2018.05.061>
- Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.
- Xu C, Lu H, Qiu P. Comparison of cell type annotation algorithms for revealing immune response of COVID-19. *Front Syst Biol* 2022;2:1–8.
- Zheng Y, Jun S-H, Tian Y *et al.* ADTnorm: Robust Integration of Single-Cell Protein Measurement across CITE-seq Datasets. *Res Sq [Preprint]*. 2024 Jul 8;rs.3.rs-4572811. <https://doi.org/10.21203/rs.3.rs-4572811/v1>
- Ziegenhain C, Vieth B, Parekh S *et al.* Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;65:631–43.e4. <https://doi.org/10.1016/j.molcel.2017.01.023>