




1 scRNAClust: A Clustering Pipeline For mammalian brain 2 single-cell RNA-sequencing data

3 **Emiliano Penalzoza^{*1}, Jiayue Tian¹, Bazillah Zargar¹, Fatemeh**
4 **Gholizadeh¹, and Camila P. E. de Souza¹**

5 ¹ University of Western Ontario

DOI: [10.21105/joss.0XXXX](https://doi.org/10.21105/joss.0XXXX)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Editor Name](#) 

Submitted: 01 January XXXX

Published: 01 January XXXX

License

Authors of papers retain
copyright and release the work
under a Creative Commons
Attribution 4.0 International
License ([CC BY 4.0](#)).

6 Summary

7 Unsupervised clustering is the primary way of identifying cell subpopulation groups in single-cell
8 RNA-sequencing (RNA-seq) data sets. The presented pipeline, scRNAClust, uses Python and
9 R and provides a lightweight interface for preprocessing, clustering, and analyzing mammalian
10 brain single-cell RNA-seq data. scRNAClust includes four state-of-the-art single-cell RNA-seq
11 clustering algorithms, Giniclust([Lan et al. \(2016\)](#)), Sc3([Kiselev et al. \(2016\)](#)), Seurat([Satija](#)
12 [et al. \(2015\)](#)) and Backspin([Zeisel et al. \(2015\)](#)). This easy-to-use pipeline intakes single-cell
13 RNA-seq data through a preprocessing step in which data are prepared into various formats
14 digestible by each of the stated clustering techniques. The following clustering performance
15 evaluation metrics are implemented: cluster purity, V-measure ([Rosenberg et al. \(2007\)](#)) and
16 ARI ([Hubert & Arabie \(1985\)](#)). Furthermore, visualization of evaluation criteria and a visual
17 representation of cluster groups are produced and easily retrieved for further analysis. This
18 pipeline simply executed through Snakemake's command-line interface provides users with the
19 necessary tools to produce powerful analysis in a swift matter. Due to this being a popular field
20 of research, the pipeline supports the integration of new clustering algorithms. The software
21 is easily accessible through its GitHub repository and can be executed on any local or remote
22 workspace.

23 Statement of Need

24 The mammalian brain is a complex system composed of specialized cells that vary in morphol-
25 ogy and functionality. Distinct cell types in the brain play different and specialized roles in
26 electrical signaling, metabolic coupling, axonal unsheathing, regulation of blood flow, and im-
27 mune surveillance ([Lou & Callaway \(2008\)](#)). Only five primary cell types have been identified:
28 neurons, astrocytes, oligodendrocytes, microglia, and endothelial cells, which are believed to
29 be responsible for the outlined functions. However, depending on the source, roughly 20-50
30 cell types have been identified as separate entities ([McKenzie et al. \(2018\)](#)). Over the past
31 few years, a series of comprehensive RNA-seq experiments in different brain cell types have
32 been published in humans and mice to try and more concretely outline the cell landscape of
33 the brain. One of the main interests in single-cell RNA sequencing is clustering, which enables
34 the extraction of the underlying subpopulations of various cell groups. By considering the
35 transcriptome of each cell, single-cell RNA-seq clustering can generate a high-resolution view
36 of gene expression in cell populations ([Ji & Ji \(2016\)](#)). Often the compiled datasets present
37 challenges as they are often zero-inflated due to specific lab protocols, as well as they have a
38 higher quantity of genes than the number of cells. Unfortunately, the described characteristics

^{*}first author

are not optimal for typical clustering algorithms. As such, specialized single-cell RNA seq clustering algorithms have been engineered to compensate for these data characteristics (Ciortan & Defrance (2021)). This introduces the need to evaluate such methods and compare them against each other. Many pipelines and workflows have been introduced to tackle this specific issue, but there is little work done addressing the need to evaluate methods for clustering brain cells based on their transcriptome profiles. Generating a more concrete picture of the brain's landscape can be crucial in solving issues related to learning, memory, and other cognitive functions. The presented software enables researchers to address the outlined problems and more efficiently conduct research

References

- Ciortan, M., & Defrance, M. (2021). Contrastive self-supervised clustering of scRNA-seq data. *NCBI*. <https://doi.org/10.1186/s12859-021-04210-8>
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*.
- Ji, Z., & Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13). <https://doi.org/10.1093/nar/gkw430>
- Kiselev, V. Y., Kirshner, K., & et al. (2016). Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*.
- Lan, J., Chen, H., Pinello, L., & Yuan, geo-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biology*.
- Lou, L., & Callaway, E. M. (2008). Genetic dissection of neural circuits. *Nature*. <https://doi.org/10.1016/j.neuron.2008.01.002>
- McKenzie, A. T., Wang, M., Hauberg, M. E., & et al. (2018). Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Nature*. <https://doi.org/10.1038/s41598-018-27293-5>
- Rosenberg, A., Hirshberg, J., & Schier, A. F. (2007). *V-measure: A conditional entropy-based external cluster evaluation measure*.
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*.
- Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., & et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*.