nature portfolio

https://doi.org/10.1038/s41596-021-00534-0

Supplementary information

Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods

In the format provided by the authors and unedited

Supplementary tables

Supplementary Table 1: Summary of annotation tools

Tool	Type	Language	Resolution	Approach	Allows "None"	Notes
clustifyr[1]	Reference-based	R	Clusters of cells	Maximum correlation	No	No statistical test, no benchmark available, demonstrated on reference bulk-RNAseq and microarray profiles
singleCellNet[2]	Reference-based	R	Single cells	Relative-expression gene-pairs + Random Forest	Yes, but rarely does so even when it should[3]	10X-100X slower than other methods. High accuracy.
scID[4]	Reference-based	R	Single cells	Discriminant analysis + gaussian mixture models	No	Only one premade reference for the mouse brain.
scmap-cluster[5]	Reference-based	R	Single cells	Consistent correlations	Yes	Fastest method available. Balances false-positives and false-negatives. Includes web-interface for use with a large set of pre-built references or custom reference.
scmap-cell[5]	Reference-based	R	Single cells	Approximate nearest neighbours	Yes	Assigns individual cells to nearest neighbour cells in reference; allows mapping cell trajectories. Fast and scalable.
singleR[6]	Reference-based	R	Single cells	Hierarchical clustering, Spearman correlations	No	Large pre-collated reference. Does not scale to data sets of 10,000 cells or more. Includes web-interface with pre-built reference
scMatch[7]	Reference-based	python	Single cells	Correlation similarity	No	Scalable to large data sets. Reference data sets combining those from SingleR and FANTOM5. Assigns standardized cell ontology cell-type labels. Supports hierarchical cell-type labels.
scPred[8]	Reference-based	R	Single cells	Principal component analysis, support vector machine	Yes	Can combine multiple reference data sets but these must be integrated a priori.
LAmbDA[9]	Reference-based	python	Single cells	Neural Network	No	GPU optimized, many hyper parameters. No pre-trained models available, users must perform their own training.
MetaNeighbor[10]	Reference-based	R	Single cells	Spearman correlation	No	Weighted voting of neighbouring cells.

CaSTLe[11]	Reference-based	R	Single cells	Mutual Information, Random Forest	No	Scalable to large data sets.
CHETAH[12]	Reference-based	R	Single cells	Hierarchical clustering, correlations	Yes	Allows intermediate annotations. Includes shiny graphical interface.
Scikit learn[13]	Reference-based	python	Multiple possible	KNN, SVM, RF, NMC, LDA	Depends on design	Expertise required for correct design and appropriate training of classifier while avoiding overfitting.
ACTINN[14]	Reference-based	python	Single cells	Neural network	No	Many hyperparameters that can affect accuracy. Pretrained models only available for Tabula Muris and human PBMC data.
Moana[15]	Reference-based	python	Single cells	Support Vector Machine (SVM), knn-smoothing	No	Training of the classifier may require subsetting the reference data.
Cell BLAST[16]	Reference-based	python	Single cells	Neural network, low dimensional embedding	Yes	Finds similar cells to each input cell, enabling mapping of cell trajectories. Web-server with pre-trained model applicable to multiple species.
OnClass[17]	Reference-based	python	Single cells	Non-linear projection	Yes	Cells are projected onto a cell-ontology space learned from a reference data set. Uses cell-ontology hierarchy to extend general labels beyond specific cell-types present in the training data. Pretrained model for the Tabula Muris data available. Once trained, the model is scalable to millions of cells.
AUCell[18]	Marker-based	R	Single cells	Area Under the Curve to estimate enrichment of markers	Yes	Due to low detection rates at the level of single cells requires many markers for every cell-type.
SCINA[19]	Marker-based	R	Single cells	Expectation-Maximization, Gaussian mixture model	(optional)	Simultaneously clusters and annotates cells. Robust to the inclusion of incorrect marker genes.
SCSA[20]	Marker-based	python	Clusters of cells	Evidence-weighted scores	Yes	Uses integrated reference from CellMarker and CancerSEA for human and mouse but allows custom marker tables. Automatically reformats outputs from scanpy, scran, Cellranger or Seurat.
Digital Cell Sorter[21]	Marker-based	python	Clusters of cells	Weighted sums of marker gene expression	Yes	Includes a clustering pipeline. Supports markers that are shared across multiple cell-types. Permutation-based significance calculation.

scCATCH[22]	Marker-based	R	Clusters of cells	Wilcoxon rank sum test, then Evidence-weighted scores	Yes	Provides integrated reference from: CancerSEA, CellMarker, MCA, CD Marker Handbook for human and mouse. Built-in identification of cluster-marker genes. Compatible with Seurat analysis pipeline.
GSEA/ GSVA[23,24]	Marker-based	R/Java	Clusters of cells	Enrichment test	Yes	Markers must all be differentially expressed in the same direction in the cluster.
Garnett[25]	Reference & Marker-based	R	Single cells or clusters of cells	Elastic-net regression classifier	Yes	Uses known markers to train a classifier on a reference data set which then annotates a novel data set. Supports hierarchical cell-type labels. Both positive and negative markers are usable. Pre-trained classifiers for human lung, mouse lung, human pbmcs, mouse brain, and C. elegans available.
scANVI/ scVI[26]	Integration	python	Single cells	Neutral network	Yes	Integrates data by estimating a common generative model across samples. Scales to large data sets.
Harmony[27]	Integration	R	Single cells	Iterative clustering and adjustment	Yes	Integrates only lower-dimensional projections of the data. Seamlessly integrated into Seurat pipeline.
Seurat-CCA[28]	Integration	R	Single cells	MNN-anchors + Canonical correlation analysis	Yes	Accuracy depends on accuracy of MNN-anchors.
mnnCorrect[29]	Integration	R	Single cells	MNN-pairs + SVD	Yes	Accuracy depends on accuracy of MNN-pairs.
LIGER[30]	Integration	R	Single cells	Non-negative matrix factorization	Yes	Allows interpretation of data set specific and shared factors of variation

** Acronyms: MNN = mutual-nearest neighbour, SVD = singular value decomposition, SVM = support vector machine, RF = random forest, KNN = k-nearest-neighbours, LDA = Linear Discriminant Analysis, NMC = nearest-mean classifier

Supplementary References

1. Fu, R. *et al.* clustifyr: an R package for automated single-cell RNA sequencing cluster classification. *F1000Res.* **9,** 223 (2020).
2. Tan, Y. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst.* **9,** 207-213.e2 (2019).
3. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20,** 194 (2019).

4. Boufea, K., Seth, S. & Batada, N. N. Mapping transcriptionally equivalent populations across single cell RNA-seq datasets. *BioRxiv* (2018). doi:10.1101/470203
5. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15,** 359–362 (2018).
6. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20,** 163–172 (2019).
7. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35,** 4688–4695 (2019).
8. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **20,** 264 (2019).
9. Johnson, T. S. *et al.* LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics* **35,** 4696–4706 (2019).
10. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9,** 884 (2018).
11. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE* **13,** e0205499 (2018).
12. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **47,** e95 (2019).
13. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12,** 2825–2830 (2011).
14. Ma, F. & Pellegrini, M. ACTINN: automated identification of cell types in single cell RNA

sequencing. *Bioinformatics* **36,** 533–538 (2020).

15. Wagner, F. & Yanai, I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *BioRxiv* (2018). doi:10.1101/456129
16. Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C. & Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat. Commun.* **11,** 3458 (2020).
17. Wang, S., Pisco, A. O., Karkanias, J. & Altman, R. B. Unifying single-cell annotations based on the Cell Ontology. *BioRxiv* (2019). doi:10.1101/810234
18. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15,** 2247–2276 (2020).
19. Zhang, Z. *et al.* SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel)* **10,** (2019).
20. Cao, Y., Wang, X. & Peng, G. SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. *Front. Genet.* **11,** 490 (2020).
21. Domanskyi, S., Hakansson, A., Bertus, T., Paternostro, G. & Piermarocchi, C. Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics. *BioRxiv* (2020). doi:10.1101/2020.07.17.208710
22. Shao, X. *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23,** 100882 (2020).
23. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23,** 3251–3253 (2007).
24. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
25. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16,** 983–986 (2019).
26. Xu, C. *et al.* Harmonization and Annotation of Single-cell Transcriptomics data with Deep Generative Models. *BioRxiv* (2019). doi:10.1101/532895

27. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16,** 1289–1296 (2019).
28. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
29. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).
30. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177,** 1873-1887.e17 (2019).