Systematic Literature Review and Meta-Analysis

Journal Homepage: https://slr-m.com/index.php/home

Research Article

Automatic cell type annotation using supervised classification: A systematic literature review

Nazifa Tasnim Hia[1]*, Sumon Ahmed[2]

[1,2]Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

*Corresponding Author email: bsse0930@iit.du.ac.bd

Submitted: 01 June 2022 *Revised: 25 September 2022* *Accepted: 30 September 2022*

ABSTRACT

Single-cell sequencing gives us the opportunity to analyze cells on an individual level rather than at a population level. There are different types of sequencing based on the stage and portion of the cell from where the data are collected. Among those Single Cell RNA seq is most widely used and most application of cell type annotation has been on Single-cell RNA seq data. Tools have been developed for automatic cell type annotation as manual annotation of cell type is time-consuming and partially subjective. There are mainly three strategies to associate cell type with gene expression profiles of single cell by using marker genes databases, correlating expression data, transferring levels by supervised classification. In this SLR, we present a comprehensive evaluation of the available tools and the underlying approaches to perform automated cell type annotations on scRNA-seq data.

Keywords: *Automatic Cell Type Annotation, Supervised Classification, Systematic Literature Review*

1. INTRODUCTION

In the central dogma of biology, it's known that DNA is transcribed into pre-mRNA, then it's processed into mature mRNA than translated into protein (Fig. 1). For the process of RNA sequencing (RNA-Seq) RNA of tissues are isolated and a snapshot is taken of them in the exact point before the translation to protein. As we all know that all cell of our body has the same DNA but they eventually became different cells because they express different genes.

So, for knowing what are the different genes that are being expressed RNA sequencing is done. Still the expressions got by the sequencing is not clear enough as it was done on the RNA of a tissue. A tissue contains around 30 trillion cells. So, it's quite impossible to understand the underlying cells behavior. As it gives us an average expression level of all the cells that the tissue contains.

https://doi.org/10.54480/slrm.v3i3.45

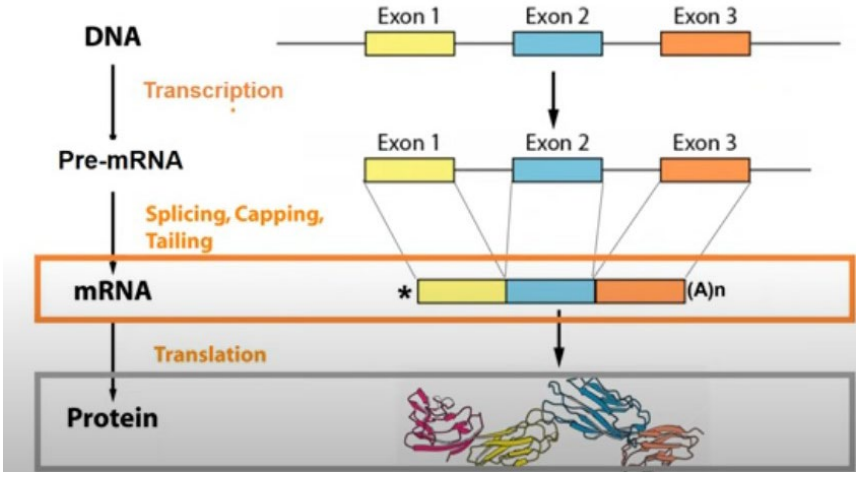

Fig. 1. Brief description of RNA sequencing.

For solving this issue, a new technology has emerged that is Single-cell RNA sequencing (scRNA-Seq). scRNA-Seq in particular, uses optimized next-generation sequencing technologies to analyze individual cells, leading to a better understanding of cell function at the genetic and cellular levels (Pasquini et al., 2021b). It allows researchers to analyze cellular heterogeneity and transcriptome heterogeneity at the single-cell level (Qi et al., 2021).

As we have the cell level data, now comes the most important job that is to identify the cell type based on the captured gene expression from scRNA-seq. However, annotating cell types by hand takes a long time and has a low level of repeatability. Computational methods for the automated annotation of cell have recently evolved to overcome these constraints. In brief, there are mainly three different techniques to identify cells automatically. Those are Correlation Based, Marker Genes Based and Supervised Classification. We are going to focus on the Supervised Classification based techniques in this review.

2. METHODOLOGY:

This review is conducted by following the protocol written in Systematic Literature Review (SLR) (Kitchenham et al., 2007). Review needs to follow three main phases which is the protocol of SLR. First phase is selecting the search strings or search keywords based on the research topic and deciding on the selection criteria. Second phase is running the search on the databases using the keywords and retrieving the studies selected based on the selection criteria. Final phase is analyzing and synthesizing the extracted papers and presenting them.

2.1. RESEARCH QUESTION:

Below are the Research Questions this review paper is addressing:

RQ1: Is feature selection helpful in the automatic cell type annotation when the approach follows supervised classification?

RQ2: What are the computational approaches in annotating cell?

2.2. Search Strategy:

At First, three standard databases are selected (See Table 1). Then based on the research topic and Research question some keywords are detected. Those are "Automated Cell Type Identification", "Single Cell RNA-Seq Data", "Supervised Classification". Complex search strings were generated using these keywords with logical operators (See Table 1). Then the search was conducted using the range 2018-2022. I have used Publish or Perish for searching from the Databases Google Scholar, Scopus . Searched directly from the Database Science Direct as it's not available from Publish or Perish.

Table 1. Search strings with condition.

Database Name	Search String	Condition
Google Scholar	("Automated Cell Type Identification" OR "Automated Cell Type Annotation") AND "Single Cell RNA Seq" AND "Supervised Classification"	2018- 2022
Science Direct	("Automated Cell Type Identification" OR "Automated Cell Type Annotation") AND "Single Cell RNA Seq"	2018- 2022
Scopus	("Automated Cell Type Identification" OR "Automated Cell Type Annotation") AND "Single Cell RNA Seq"	2018- 2022

2.3. Selection Criteria:

Through searching relevant and irrelevant papers both are retrieved. A selection criterion is needed for selecting the relevant ones from the search result. This selection criteria are divided in three steps. At first need to identify the duplicate ones and remove them as search is conducted on three different databases and sometimes same database search also returns duplicate papers. Also need to identify the non-English language papers and exclude them. The paper found after this step are ready for going through screening based on title and abstract. Papers will be excluded from this step if the abstract and title seems unaligned with the research interest. The papers retrieved after this step are ready for full text analysis. If any paper doesn't have full text access that will be excluded.

2.4. Studies Selection:

34 records were selected through the strategic search compiled on three databases. Following the flow of PRISMA (Kitchenham et al., 2007) (See Fig. 2) at first the duplicates found between and within the search results of three Databases were removed. There were 5 duplicates after removing those we got 29 papers for further inquiry. Next elimination round was based on the title and abstract part of the papers. After reviewing the title and abstract part 8 results were found irrelevant. After removing those, remaining 21 papers were selected for full text analysis and full text of all of them were available.

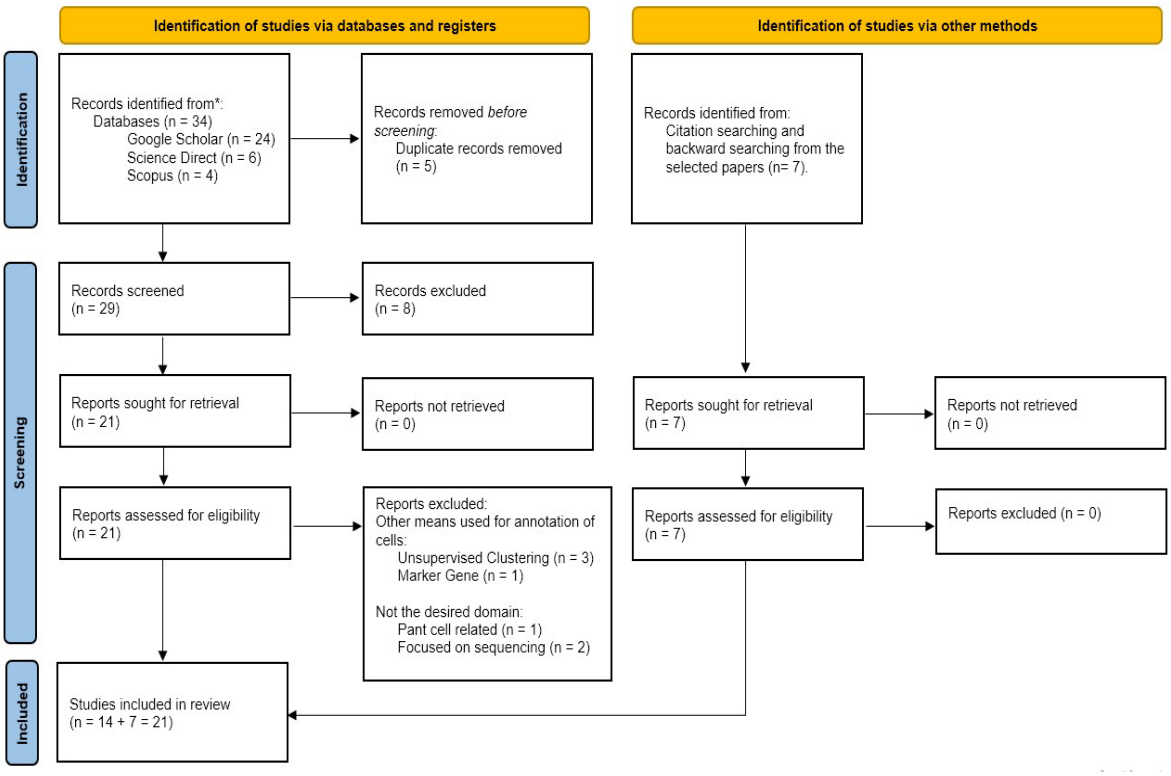

Fig. 2. Flow of searching to selection using PRISMA flowchart.

7 of them were found not to satisfy the research questions. As the first RQ was looking for the supervised classification approach of annotating cell automatically. Two of them were found using unsupervised clustering and one of them using Marker gene-based classification. As the research topic is for annotation of single cell of human or animal. There was one focused on plant cell and another one was focused on sequencing process of single cell. By these, it came to 14 papers from 34.

There are 4 review papers included in these 14 papers. Seven more papers were identified with the help of forward and backward referencing on these review papers and the others. Accumulating these papers with the previous ones finally it became 21 records for result analysis.

3. DATA EXTRACTION, SYNTHESIS AND REPORTING:

Extracted papers based on selection criteria are analyzed here in this section. Here descriptive analysis and the findings of review will be reported

3.1. Result and Discussion:

Results found from databases search and forward backward referencing were filtered and ultimately came to 21 selected papers. This section contains analysis based on the details of those papers such as published year, source database, Languages used in the computational approaches.

3.1.1. Publications Based on Year:

The Fig. 3 shows the number of publications published per year between 2018-2022 which are under review. From this graph it's interpretable that the number of automatic annotation tools those used supervised classification as approach are increasing in number. As the review is being conducted in the first quarter of the 2022 so it's reasonable that 2022 has less publications then 2021.

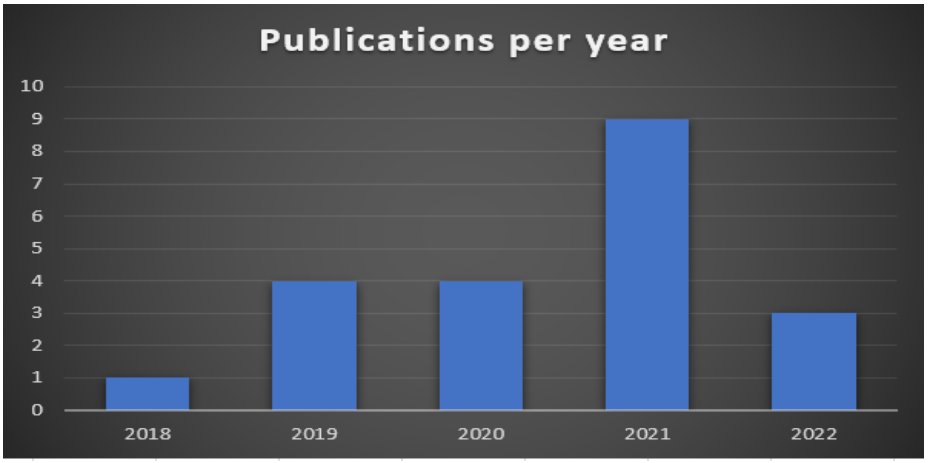

Fig. 3. Publications distributed over the years.

3.1.2. Publications Based on Database:

Three Databases are used for this review are Google Scholar, Scopus and Science Direct. 72% of the selected findings are from google scholar, 21% are from Scopus and 7% are from science direct (See Fig. 4). Thus, it can be said that google scholar is the prime source of the findings for this review.

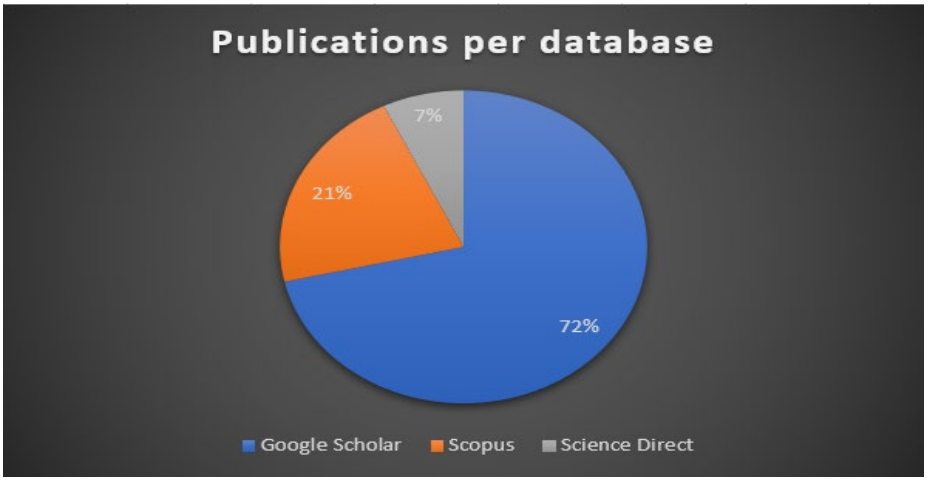

Fig. 4. Publications distributed over the databases.

3.1.3. Languages used in developing the computational approaches of the tools:

As all the computational approaches are in the Machine Learning domain. So, it doesn't have many options anyways. Python and R are known as the best ones for this domain.

Table 2. Languages used in the computational approaches.

Tool Name	Language Used	Ref
Moana	Python	(Wagner & Yanai, 2018)
LAmbDA	Python	(Johnson et al., 2019)
superCT	Python	(Xie et al., 2019)
SingleCellNet	R	(Tan & Cahan, 2019)
Garnet	R	(Pliner et al., 2019)
scPred	R	(Alquicira-Hernandez et al., 2019)
ACTINN	Python	(Ma & Pellegrini, 2020)
OnClass	Python	(Wang et al., 2019)
scClassify	R	(Lin et al., 2020)
scANVI	Python	(Xu et al., 2021)
scNym	Python	(Xu et al., 2021)
Superscan	Python	(Shasha et al., n.d.)
scAnnonate	R	(Ji et al., 2022)
MarkerCapsule	Python	(Ray & Schönhuth, 2020)
Besca	Python	(Mädler et al., 2021)
RMTL	Python	(Upadhyay et al., n.d.)
JIND	Python	(Goyal et al., 2022)

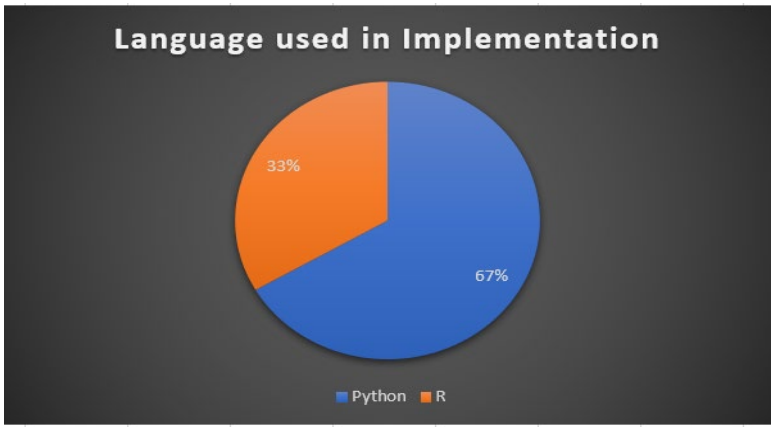

Fig. 5. Languages Used for the computational approaches of the publications in Percentage.

From the Fig. 5 and Table 2, it shows no exception than that. 67% of the approaches are written in Python and the rest 33% uses. It can be said that Python is more popular among these two. Though it totally depends on the programmer preference. As both of them offer more or less same flexibilities.

4. RELATING RESEARCH QUESTIONS TO THE SELECTED PUBLICATIONS:

4.1. Is feature selection helpful in the automatic cell type annotation when the approach follows supervised classification?

Feature selection is a process to reduce the input variable based on the relevancy of the variable with the goal. It helps to increase the quality of the data and ease the analysis as it reduces the data size. Though with present technologies it's not a problem to work with the standard scRNA-seq dataset sizes. However, if the data have unnecessary information (features) it can be misleading.

As single cell data could have many zero reads of the genes expression levels and all genes are not important for cell type identification. So, we can say single cell data is quite noisy without proper processing or filtration may lead the classification model to cause overfit and thus suboptimal performances.

Almost all the tools under this review either have a pre-processing step for their data or runs feature selection process before proceeding to training the model.

SingleCellNet (Tan & Cahan, 2019), ACTINN (Ma & Pellegrini, 2020), Moana (Wagner & Yanai, 2018), scPred (Alquicira-Hernandez et al., 2019) are four tools under this review which use feature selection prior to training. An experiment was done in this (Theunissen, 2021) review paper which concludes that feature selection leads to classification improvement. “If datasets with meaningful features and sufficient label representation are available, supervised learning methods might offer a powerful and flexible alternative for their analysis” (Pasquini et al., 2021b).

It summarizes that feature selection can help in a positive way for the automation of cell type and it’s an important step.

4.2. What are the computational approaches in annotating cell?

Automatic cell type annotation approaches try to find commonalities between scRNA-seq datasets while accounting for the data's inherent noise and variability. Indeed, the variability reported between scRNA-seq datasets is due to a number of confounding factors. Machine learning approaches have proven to be an excellent resource for a range of tasks in analysis pipelines, including dimensionality reduction operations, due to the characteristic noise and multidimensionality of scRNA-seq data. Supervised classification, or the transfer of labels from labeled to unlabeled datasets, is a classic machine learning paradigm for which many techniques have been developed. The term 'supervised learning' is used in the field of machine learning to describe the construction of a model distribution of labels (cell types) in terms of a set of features (genes) that is trained on ground truth data (a previously annotated dataset).

Table 3. Computational approaches used in the tools.

Tool Name	Computational Approach	Ref
Moana	kNN-smoothing + SVM	(Wagner & Yanai, 2018)
LAmbDA	Multiple ML Techniques	(Johnson et al., 2019)
superCT	Artificial Neural Network	(Xie et al., 2019)
SingleCellNet	Random Forest	(Tan & Cahan, 2019)
Garnet	Elastic net regression	(Pliner et al., 2019)
scPred	SVM	(Alquicira-Hernandez et al., 2019)
ACTINN	Artificial Neural Network	(Ma & Pellegrini, 2020)
OnClass	kNN and Bilinear Neural Network	(Wang et al., 2019)
scClassify	Weighted kNN Classifier	(Lin et al., 2020)
scANVI	kNN Classifier	(Xu et al., 2021)
scNym	Adversarial Neural Network	(Xu et al., 2021)
Superscan	XGBoost (eXtreme Gradient Boosting)	(Shasha et al., n.d.)
scAnnonate	Marginal model-based ensemble learning	(Ji et al., 2022)
MarkerCapsule	Supervised learning with capsule network	(Ray & Schönhuth, 2020)

Tool Name	Computational Approach	Ref
Besca	SVM and logistic regression	(Mädler et al., 2021)
RMTL	Regularized multi-task learning	(Upadhyay et al., n.d.)
JIND	Artificial Neural Network	(Goyal et al., 2022)

Though approach of automatic annotation of cell can be divided into main three branches such as Correlation based, Marker Gene based and Supervised Classification. However, only supervised classification related approaches are listed above as this review is focused on the supervised classification related approaches. Almost all possible methodologies of supervised classification are present in the list such as neural networks, support vector machines, ensemble learning etc. From the above list (Table 3) it's not viable to say that any one of the approaches is the most effective approach. Though Neural network and KNN seem the most applied approach among these 17 approaches.

5. CONCLUSION AND LIMITATIONS:

A systematic Literature Review on Automatic Cell Annotation using Supervised Classification has been done. 21 papers are selected after all searching, forward backward referencing and filtering following the SLR principles (Kitchenham et al., 2007). In this review some descriptive analysis and answers of research questions are provided which are taken from the selected papers. These will be very helpful for anyone who is interested to work with Automatic annotation of single cell.

There are obviously some drawbacks of this review as this was a bounded work for time constraint. This review has used only three databases as source of research papers. In research there are many factors to look into but this paper only focused on the two specific factors that supported the research questions.

Nonetheless, this review is a concise and useful one for anyone who is interested in this domain.

Author Contributions:

Nazifa Tasnim Hia and Sumon Ahmed conceptualized and validated the methodology for this research. The data was analyzed by Nazifa Tasnim Hia. Both authors contributed to the final manuscript version. The published version of the manuscript has been read and approved by all authors. The study was supervised by Sumon Ahmed.

Funding:

This research received no external funding.

Data Availability Statement:

The data used in this study is publicly available. Three Databases used for this review are Google Scholar, Scopus and Science Direct.

Acknowledgments:

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-magnificent, the Ever-Thankful. We would like to express my sincere gratitude to Md. Mahbubul Alam Joarder for his

insightful and constructive suggestions during the research planning and development process. His willingness to give his time so generously has been greatly appreciated. Finally, I'd like to thank my parents for their motivation and help throughout my studies.

Conflicts of Interest:

The authors declare no conflict of studies.

Reference:

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., & Powell, J. E. (2019). ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, *20*(1), 1–17. https://doi.org/10.1186/s13059-019-1862-5

Goyal, M., Serrano, G., Argemi, J., Shomorony, I., Hernaez, M., & Ochoa, I. (2022). JIND: joint integration and discrimination for automated single-cell annotation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btac140

Ji, X., Tsao, D., Bai, K., Tsao, M., & Zhang, X. (2022). scAnnotate: an automated cell type annotation tool for single-cell RNA-sequencing data. *BioRxiv*, 2022.02.19.481159. https://www.biorxiv.org/content/10.1101/2022.02.19.481159.abstract

Johnson, T. S., Wang, T., Huang, Z., Yu, C. Y., Wu, Y., Han, Y., Zhang, Y., Huang, K., & Zhang, J. (2019). LAmbDA: Label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics*, *35*(22), 4696–4706. https://doi.org/10.1093/bioinformatics/btz295

Kitchenham, B. A., Mendes, E., & Travassos, G. H. (2007). Cross versus within-company cost estimation studies: A systematic review. *IEEE Transactions on Software Engineering*, *33*(5), 316–329. https://doi.org/10.1109/TSE.2007.1001

Lin, Y., Cao, Y., Kim, H. J., Salim, A., Speed, T. P., Lin, D. M., Yang, P., & Yang, J. Y. H. (2020). scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular Systems Biology*, *16*(6). https://doi.org/10.15252/MSB.20199389

Ma, F., & Pellegrini, M. (2020). ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics*, *36*(2), 533–538. https://doi.org/10.1093/bioinformatics/btz592

Mädler, S. C., Julien-Laferriere, A., Wyss, L., Phan, M., Sonrel, A., Kang, A. S. W., Ulrich, E., Schmucki, R., Zhang, J. D., Ebeling, M., Badi, L., Kam-Thong, T., Schwalie, P. C., & Hatje, K. (2021). Besca, a single-cell transcriptomics analysis toolkit to accelerate translational research. *NAR Genomics and Bioinformatics*, *3*(4). https://doi.org/10.1093/nargab/lqab102

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021a). Automated methods for cell type annotation on scRNA-seq data. In *Computational and Structural Biotechnology Journal* (Vol. 19, pp. 961–969). https://doi.org/10.1016/j.csbj.2021.01.015

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021b). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, *19*, 961–969. https://doi.org/10.1016/j.csbj.2021.01.015

Pliner, H. A., Shendure, J., & Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nature Methods*, *16*(10), 983–986. https://doi.org/10.1038/s41592-019-0535-3

Qi, R., Wu, J., Guo, F., Xu, L., & Zou, Q. (2021). A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/bib/bbaa216

Ray, S., & Schönhuth, A. (2020). MarkerCapsule: Explainable Single Cell Typing using Capsule Networks. In *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.09.22.307512.abstract

Shasha, C., Tian, Y., Mair, F., Miller, H., BioRxiv, R. G.-, & 2021, U. (n.d.). Superscan: Supervised Single-Cell Annotation. *Biorxiv.Org*. Retrieved May 12, 2022, from https://www.biorxiv.org/content/10.1101/2021.05.20.445014.abstract

Tan, Y., & Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, *9*(2), 207-213.e2. https://doi.org/10.1016/j.cels.2019.06.004

Theunissen, L. (2021). *A COMPARISON OF FLAT AND HIERARCHICAL CLASSIFICATION FOR AUTOMATIC ANNOTATION OF SINGLE-CELL TRANSCRIPTOMICS DATA*. https://libstore.ugent.be/fulltxt/RUG01/003/008/162/RUG01-003008162_2021_0001_AC.pdf

Upadhyay, P., Genetics, S. R.-F. in, & 2022, U. (n.d.). A Regularized Multi-Task Learning Approach for Cell Type Detection in Single-Cell RNA Sequencing Data. *Europepmc.Org*. Retrieved May 12, 2022, from https://europepmc.org/articles/pmc9043858/bin/datasheet1.pdf

Wagner, F., & Yanai, I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *BioRxiv*. https://doi.org/10.1101/456129

Wang, S., Pisco, A. O., McGeever, A., Brbic, M., Zitnik, M., Darmanis, S., Leskovec, J., Karkanias, J., & Altman, R. (2019). Unifying single-cell annotations based on the Cell Ontology. *BioRxiv*, 810234. https://doi.org/10.1101/810234

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., Von Hoff, D., Han, H., Zhang, M. Q., & Lin, W. (2019). SuperCT: A supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Research*, *47*(8), 1–12. https://doi.org/10.1093/nar/gkz116

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., & Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, *17*(1). https://doi.org/10.15252/msb.20209620