Marta Interlandi* and Martin Dugas

Cell type classification in single-cell RNA sequencing: Review of methods

Institut für Medizinische Informatik
IMI
Albert-Schweitzer-Campus 1, 48149 Münster
*email: Marta.Interlandi@ukmuenster.de

Introduction

Traditionally, the definition of a cell type has been performed considering unique cellular shapes or structures, the expression of specific marker proteins or the anatomic location. With the advent of single-cell RNA sequencing (scRNA-seq), it is possible now to extract information on the expression of thousands of genes in thousands of cells comprising the tissue of interest. This enables the identification of subpopulations of cells that share a common pattern of gene expression, typically through dimensionality reduction techniques (PCA, t-SNE) and clustering. One of the goals is then to classify each subpopulation according to the corresponding cell type. Here we present the first results of a review of methods employed, until now, for cell type classification. We could delineate three major approaches: (i) Empirical methods based on known makers of cell types that require the manual classification by an expert; (ii) Semi-automated methods that rely on known markers but try to combine them with machine learning techniques; (iii) Fully automated methods using machine learning (i.e. neural networks) and collections of published single-cell classified datasets.

Methods

We collected 16 papers by searching in literature for "cell type classification", "cell type discovery" and by manual selection of related papers (see Table 1, attached).

Cell type classification

Empirical methods - supervised

These methods are generally the most widely used in single-cell studies. We could delineate two main complementary approaches:

1 For known or expected cell types, the method consists in defining a set of marker genes and manually classifying the clusters of cells (*Tirosh et al.*, *Zheng et al.*). At this purpose, *Danaher et al.* identified a list of 60 markers for 14 immune cell types.

2 Starting with the identification of cell clusters and the calculation of differentially expressed genes as markers of each cluster, the classification relies on a functional annotation of those markers (*Usoskin et al.*, *Campbell et al.*).

Semi-supervised methods

In this case the classification is still based on markers of subpopulations, but there is an attempt of automating the process. *Schelker et al.* identified a set of 45 markers divided in 3 categories (AND/OR/NOT genes) and then performed the classification using a decision tree. *Scialdone et al.* compared 5 established machine learning methods and a custom-build predictor for the classification of cells based on their cell-cycle phase. *Aevermann et al.* used random forests to identify a set of necessary and sufficient markers to be used in the definition of consistent and reproducible cell types in the context of the Cell Ontology (CL) (*Bard et al.*).

Unsupervised methods

To this category belong those approaches that, instead of relying on sets of markers, gather an extensive collection of datasets and use them to train a machine learning algorithm (i.e. neural networks (NN)). *Lin et al.* collected 33 datasets, tested various NN architectures and built a web interface where it is possible to upload a dataset (TPM values, max size = 200 MB) and receive an email with the classification in terms of 100 nearest neighbors per cell. *Alavi et al.* applied a similar approach to over 500 different studies with more than 300 unique cell types. The analysis is however limited to mouse data and the classification can only be visualized in the web server. *Kiselev et al.* aimed at identifying cell populations by projecting them on a reference dataset where the cell identity is known.

Results

We tested the performance of supervised methods in the classification of immune cell types. We downloaded scRNA-seq data from 10X genomics webpage, choosing only the datasets comprised of a known cell type (enriched from fresh PBMCs by FACS). From a total number of 9 datasets (thus, 9 cell types) we retained 2500 cells per dataset (random sample without replacement), to have an evenly distributed ground truth population. The analysis was performed using the R package Seurat (*Butler et al.*).

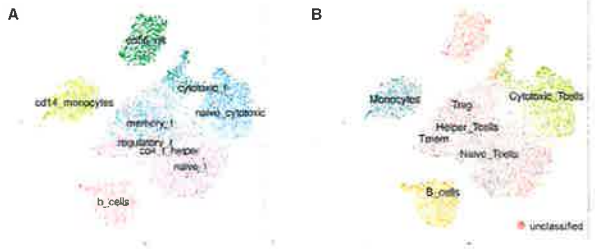
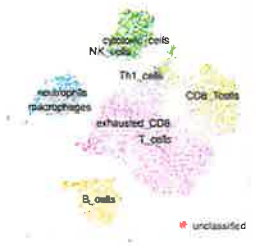


Figure A shows the t-SNE plot of the dataset labelled by the original cell types. Figure B shows the same population of cells classified using a manually curated list of markers from literature (29 markers, 9 cell types). The classification is performed considering the mean value of expression of each set of markers present in our data (assigned cell type has the highest mean value). Although correctly classifying 97% of B cells cluster, the classification failed at retrieving natural killer cells (CD56-NK) and achieved only 4% of accuracy for memory T cells and 5% for regulatory T cells.

Then, we considered a more comprehensive list of markers, choosing the set identified by *Danaher et al.* (60 markers, 14 immune cell types). In this case, as shown in Figure C, the algorithm could identify a small cluster of NK cells (4% of the initial cluster) but could not separate different T cell types in the central cluster.

These results show that the identification of different cell types in a single-cell transcriptome analysis is mostly not trivial. The task is complicated by the same nature of cells, whose boundaries to other cell types are sometimes not clearly defined, as pointed out in the article *"What is your conceptual definition of "Cell Type" in the context of a mature organism?"* by *Clevers et al.*

C



References

Aevermann, Brian D., et al. "Cell type discovery using single-cell transcriptomics: implications for ontological representation." Human molecular genetics 27.R1 (2018): R40-R47.
Alavi, Amir, et al. "scQuery: a web server for comparative analysis of single-cell RNA-seq data." *bioRxiv* (2018): 323238.
Bard, Jonathan, et al. "An ontology for cell types." *Genome biology* 6.2 (2005): R21.
Butler, Andrew, et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species." *Nature biotechnology* 36.5 (2018): 411.
Campbell, John N., et al. "A molecular census of arcuate hypothalamus and median eminence cell types." *Nature neuroscience* 20.3 (2017): 484.
Clevers, Hans, et al. "What is your conceptual definition of "cell type" in the context of a mature organism?." *Cell Systems* 4.3 (2017): 255-259.
Danaher, Patrick, et al. "Gene expression markers of tumor infiltrating leukocytes." *Journal for immunotherapy of cancer* 5.1 (2017): 18.
Kiselev, Vladimir Yu, et al. "scmap: projection of single-cell RNA-seq data across data sets." *Nature methods* 15.5 (2018): 359.
Lin, Chieh, et al. "Using neural networks for reducing the dimensions of single-cell RNA-Seq data." *Nucleic acids research* 45.17 (2017): e156-e156.
Schelker, Max, et al. "Estimation of immune cell content in tumour tissue using single-cell RNA-seq data." *Nature communications* 8.1 (2017): 2032.
Scialdone, Antonio, et al. "Computational assignment of cell-cycle stage from single-cell transcriptome data." *Methods* 85 (2015): 54-61.
Tirosh, Itay, et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq." *Science* 352.6282 (2016): 189-196.
Usoskin, Dmitry, et al. "Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing." *Nature neuroscience* 18.1 (2015): 145.
Zheng, Grace XY, et al. "Massively parallel digital transcriptional profiling of single cells." *Nature communications* 8 (2017): 14049.

UKM Universitätsklinikum Münster
WWU MÜNSTER
medizinische fakultät Westfälische Wilhelms-Universität Münster

Cell type classification in single-cell RNA sequencing: Review of methods

Authors: Marta Interlandi and Martin Dugas
Contact details: Marta.Interlandi@ukmuenster.de

Authors	Year	DOI	Short Description
Empirical methods - supervised			
Campbell et al.	2017	10.1038/nn.4495	scRNA seq of ~20,000 cells from adult mouse hypothalamic arcuate-median eminence complex (using Drop-seq). Indentification of 50 distinct cell populations by expression of known markers.
Chen et al.	2017	10.1038/srep45656	scRNA seq of mouse embryonic medial ganglionic eminence (MGE) (225 cells at different time points) and MGE-like cells differentiated from embryonic stem cells (113 cells). PCA, DE analysis of distinct subpopulations and annotation by Gene Ontology.
Puram et al.	2018	10.1016/j.cell.2017.10.044	scRNA seq of ~6,000 cells from 18 head and neck squamous cell carcinoma patients and 5 matched pairs of primary tumors and lymph node metastases. Identification of malignant vs. non-malignant cells, annotation of clusters by the expression of known marker genes as T cells, B/plasma cells, macrophages, dendritic cells, mast cells, endothelial cells, fibroblasts and myocytes.
Tirosh et al.	2016	10.1126/science.aad0501	scRNA seq of ~4,600 cells from 19 melanoma patients. Identification of malignant vs. non-malignant cell types (B-cells, T-cells, macrophages, endotelial cells, natural killer cells and cancer-associated fibroblasts).
Usoskin et al.	2014	10.1038/nn.3881	scRNA seq of 622 single mouse neurons. Identification of 11 distinct cell types. PCA and examination of known markers, together with DE analysis, to classify cell subpopulations.
Zheng et al.	2017	10.1038/ncomms14049	scRNA seq of 68,000 peripheral blood mononuclear cells and annotation of distinct immune cell types by the expression of known markers.
Semi-supervised methods			
Aevermann et al.	2018	10.1093/hmg/ddy100	Method based on random forest machine learning for identifying sets of necessary and sufficient marker genes, which can be used to assemble consistent and reproducible cell type definitions for incorporation into the Cell Ontology. (method: NSforest, code not publicly available)
Guo et al.	2015	10.1371/journal.pcbi.1004575	SINCERA, pipeline to analyse scRNA seq. Module "Cell type identification" performs clusters optimization, DE analysis and functional annotation (ToppGene Suite, DAVID, MSigDB and Genecards).
Schelker et al.	2017	10.1038/s41467-017-02289-3	Multi-step approach to classify scRNA-seq data. Use of a set of 45 marker genes divided in AND (all required), OR (only one required), NOT (not expressed). Initial cell type assignment to create a sparse training set, decision tree classifier trained on this training data and the identity of all cells predicted and validated based on a five-fold cross validation. (code available, Matlab)
Scialdone et al.	2015	10.1016/j.ymeth.2015.06.021	Supervised machine learning approach to evaluate the ability of 6 algorithms to predict the unobserved cell cycle stage of a cell. Methods compared are: random forest, logistic regression and lasso, support vector machines, PCA-based classification and a custom-built predictor (based on top scoring pairs classifiers). (code available)
Suffiotti et al.	2017	10.1007/s00251-017-1002-x	Approach to identify innate lymphoid cells (ILCs) using scRNA seq data. Logistic regression-based classifier trained on mouse ILC and NK gene expression data and validated in scRNA seq data of human ILCs and NK. (code not publicly available)
Unsupervised methods			
Alavi et al.	2018	10.1101/323238	Automated pipeline to download, process and annotate publicly available scRNA seq datasets (processed: 500 different studies with over 300 unique cell types, only mouse). Supervised neural network models to learn reduced dimension representations for each of the input profiles. This reduced dimension profiles are stored in a web server that allows users to perform queries to compare new data to the data collected. (Supporting website: http://scquery.cs.cmu.edu)
Kiselev et al.	2018	10.1038/nmeth.4644	scmap, a method for projecting cells from a scRNAseq dataset onto cell types or individual cells from other experiments. Identification of cell types by projecting unknow cells onto known populations. (R package)
Lin et al.	2017	10.1093/nar/gkx681	Neural networks on a training set of 33 different available datasets, implementation of a web server for cell type retrieval. Required data type is TPM, maximum upload file size is 200 MB. Results sent by email. (Supporting website: http://sb.cs.cmu.edu/scnn/)
Others			
Clevers et al.	2017	10.1016/j.cels.2017.03.006	Collection of opinions from 13 authors on the topic "definition of cell type" in a mature organism.
Danaher et al.	2017	10.1186/s40425-017-0215-8	Definition of a set of markers of tumor infiltrating leukocytes. Total list of 60 markers to identify 14 immune cell populations: B-cells, CD45, cytotoxic cells, DC, exhausted CD8, macrophages, mast cells, neutrophilis, NK CD56dim cells, NK cells, T-cells, Th1 cells, Treg, CD8 T cells.