Fully-automated cell-type identification with specific markers extracted from single-cell transcriptomic data

Aleksandr Ianevski[1,2], Anil K Giri[1*] and Tero Aittokallio[1,2,3,4,5*]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, FI-00290 Helsinki, Finland
[2]Helsinki Institute for Information Technology (HIIT), Department of Computer Science, Aalto University, FI-02150 Espoo, Finland
[3]Instute for Cancer Research, Department of Cancer Genetics, Oslo University Hospital, Norway
[4]Centre for Biostatistics and Epidemiology (OCBE), Faculty of Medicine, University of Oslo, Norway
[5]Department of Mathematics and Statistics, University of Turku, Quantum, FI-20014 Turku, Finland
*Corresponding authors

Keywords single-cell annotation, data-driven marker identification, web-application, cell marker database

Abstract

Single-cell transcriptomics enables systematic charting of cellular composition of complex tissues. Identification of cell populations often relies on unsupervised clustering of cells based on the similarity of the scRNA-seq profiles, followed by manual annotation of cell clusters using established marker genes. However, manual selection of marker genes for cell-type annotation is a laborious and error-prone task since the selected markers must be specific both to the individual cell clusters and various cell types. Here, we developed a computational method, termed ScType, which enables data-driven selection of marker genes based solely on given scRNA-seq data. Using a compendium of 7 scRNA-seq datasets from various human and mouse tissues, we demonstrate how ScType enables unbiased, accurate and fully-automated single-cell type annotation by guaranteeing the specificity of marker genes both across cell clusters and cell types. The widely-applicable method is implemented as an interactive web-tool (https://sctype.fimm.fi), connected with comprehensive database of specific markers.

Introduction

Accurate identification of distinct cell types in complex tissue samples is a critical pre-requisite for elucidating their roles in various biological processes including haematopoiesis and embryonic development[1,2]. Traditionally, cell sorting and microscopic techniques have been extensively used to isolate cell types, followed by molecular profiling of the sorted cells using, for instance, mRNA or protein measurements[3,4,5]. Decades of research has led to several collections of cell-specific

features, including expression of marker genes, that are being used to distinguish various cell types[6,7]. However, the entire process is manually tedious and technically challenging. Recently, single-cell RNA sequencing (scRNA-seq) has been established as an efficient approach to chart diverse cell populations in tissue samples and to study various biological processes in disease and development[2,8,9]. The scRNA-seq technology provides an unprecedented view of various cell types and is the leading technology in large-scale cell mapping projects such as the Human Cell Atlas.[10]

Identification of cell populations in a given sample is typically solved by unsupervised clustering of cells based on their transcriptomic profiles[11,12]. In the next step, the most differentially expressed genes between a selected cluster and all the other detected clusters are identified as marker genes. These marker genes are then manually inspected using available information in the literature or cell marker databases[6,7] to assign cell-type labels to each detected cluster. However, the manual selection of cluster-specific marker genes is a time-consuming and error-prone task, since (*i*) differentially-expressed genes are often expressed in multiple clusters, and (*ii*) the identified genes may be known markers for multiple cell-types. This manual task is further complicated by the lack of curated cell marker databases that include both known and *de novo* markers to annotate cell-types with confidence. For example, selection of CD44 as marker gene to label any single-cell type may compromise the accuracy of cell annotation as CD44 is expressed in various immune cell types.[6]

To solve these challenges, we developed a data-driven method that requires only scRNA-seq data for unsupervised selection of marker gene panels that guarantee maximal specificity across both the cell clusters and cell-types. The computational algorithm together with a comprehensive marker database enables one to identify marker genes that are uniquely expressed in any given cell cluster and are specific to a particular cell type within a tissue. The ScType platform is implemented as an open-source and interactive web-tool (https://sctype.fimm.fi), connected to a new database to enable fast, accurate and fully-automated cell-type annotation. We carried out a systematic benchmarking of ScType across 7 scRNA-seq datasets from 2 mouse and 4 human tissues, and showed that that ScType correctly annotated a total of 81 out of 82 cell-types (98.8% accuracy), including 8 newly-reannotated cell-types that were originally incorrectly or non-specifically annotated in the studies.

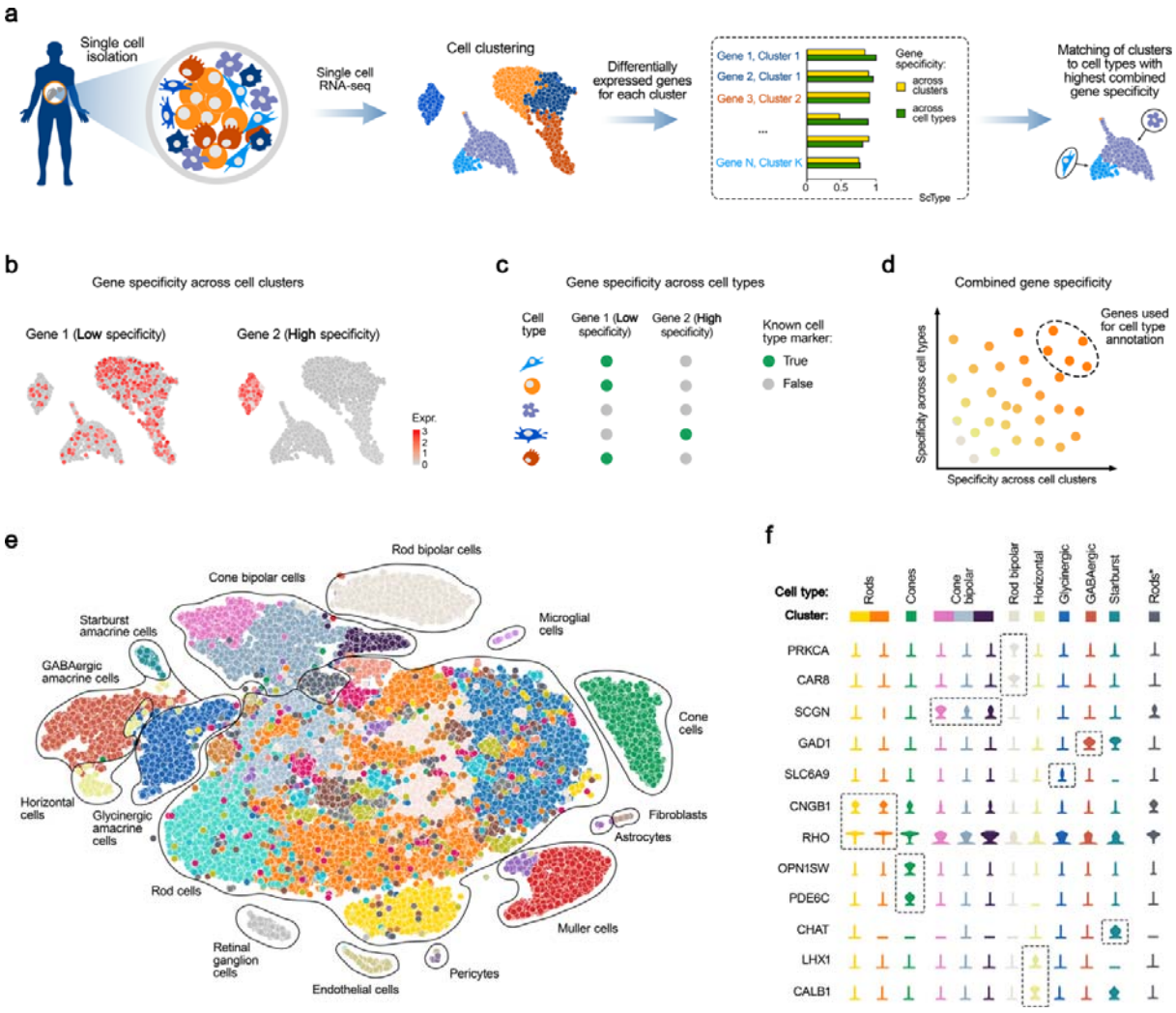


Figure 1. A schematic view of data-driven marker identification and cell-type annotation using ScType. (a) ScType requires only the raw or pre-processed single-cell transcriptomics dataset(s) as input. Additional quality control (e.g., removal of outlier cells with high mitochondrial gene expression) and normalization steps (e.g., removal of technical artefacts) are performed, where needed, and followed by unsupervised clustering of cells based on scRNA-seq profiles. ScType prioritizes markers among top positively differentially-expressed genes, according to their specificity across the clusters and cell-types. Marker genes with the highest specificity are used to label clusters using the cell-type information in the ScType database. (**b-d**) An overview of the ScType marker prioritization algorithm. The algorithm selects markers that have the highest specificity both for a given cluster (calculated using the input scRNA-seq data) and cell-type (calculated automatically using cell-type information in ScType database). (**e**) An example of cell-type labelling based on the specific markers in mouse retina extracted by ScType using scRNA-seq data from Macosko et al.[13] (**f**) Violin plots show the expression levels of selected marker genes that were used to assign the cell-type labels to each cluster. For example, ScType identifies

PRKCA[14,15] and CAR8[14] as the top specific marker genes that are uniquely expressed in the rod bipolar cells.

Results

ScType improves annotation of cell-types using solely scRNA-seq data

We first investigated the performance of ScType by re-analysing a published scRNA-seq study of mouse retinal cells[13]. ScType accurately annotated all the 12 identified retinal cell types (Fig 1e). Additionally, it automatically identified the 3 closely-related cell populations of amacrine cell types (GABAergic, glycinergic and startbust) that were originally-identified by extensive and deeper analysis of selectively-expressed markers[13], indicating that ScType enables one to accurately discriminate between cell populations with similar transcriptomic profiles. Further, ScType was able to distinguish between the subtypes of bipolar cells (rod and cone bipolar cells), that were assigned to single group in the original study, therefore enhancing the resolution of cell-type annotation.

As an example, ScType ranked *PRKCA* and *CAR8* among the top-5 marker genes specific for the rod bipolar cell (RBC) cluster (Fig. 1f); both are known RBC markers[14,15]. Three additional cell clusters were assigned to cone bipolar subtype as they uniquely-expressed *SCGN*, a well-established marker for cone bipolar cells[16]. Interestingly, ScType identified a small sub-type (<1% of cells) of rod cells that expressed *SCGN* (labelled as Rods* in Fig.1f). These cells were originally annotated as rod cells in the original manuscript[13]. These results indicate that ScType automatically prioritizes specific markers for accurate annotation of cell-types with distinct molecular features.

Systematic evaluation of ScType across multiple scRNA-seq datasets

We next benchmarked ScType performance in terms of its ability to automatically assign cell-types in comparison to the cell-type annotations given by the original authors of 7 published scRNA-seq studies. These RNA-seq datasets originated from various tissues including human liver[8], pancreas[17], peripheral blood mononuclear cells (PBMCs)[18], brain[19], mouse lung[20] and retina samples,[13] as well as a human pancreas mixture of eight previously-published datasets using tissue samples from human pancreatic islets spanning 27 donors, five technologies, and four laboratories[22] (see Methods for details). These varied scRNA-seq datasets were utilized to investigate a wider applicability of ScType to various sequencing platforms, tissues types and organisms.

ScType correctly annotated a total of 81 cell types, including 8 correctly reannotated cell-types that were originally incorrectly or non-specifically annotated (Fig. 2a). The only cell-type it was not able to label correctly was fetal cells in the human brain dataset, as there are no fetal cell markers for

human brain in the current version of the ScType database. However, ScType correctly identified other cell populations of the human brain tissues - oligodendrocytes, astrocytes, microglial cells, neurons, endothelial and oligodendrocyte precursor cells - as annotated in the original study[19]. Further, ScType was able to refine the originally-annotated neuron cell population into cholinergic (expressing SLC17A7)[22] and glutamatergic (expressing ACHE)[23] sub-types.

In the human liver dataset,[8] ScType distinguished between B cells and plasma cell-types (Fig 2b). The segregated B and plasma cells differed based on their specific expression of CD19, CD20 and CD138 markers (Fig. 2c), as $CD19^{+}$ or/and $CD20^{+}$ are uniquely expressed in B-cells whereas $CD138^{+}$ ($CD19^{-}CD20^{-}$) is uniquely expressed in plasma cells[24]. Further, ScType accurately assigned various cell types in the human pancreas dataset, where it correctly labelled the subpopulation of macrophages that was incorrectly labelled as acinar cells in the original study[17] (Supplementary Fig. 1a). The cluster- and cell type- specific marker genes identified by ScType included, among others, APOC1, C1QA, CD52 and MSR1, which are known canonical macrophage markers.[25-28]

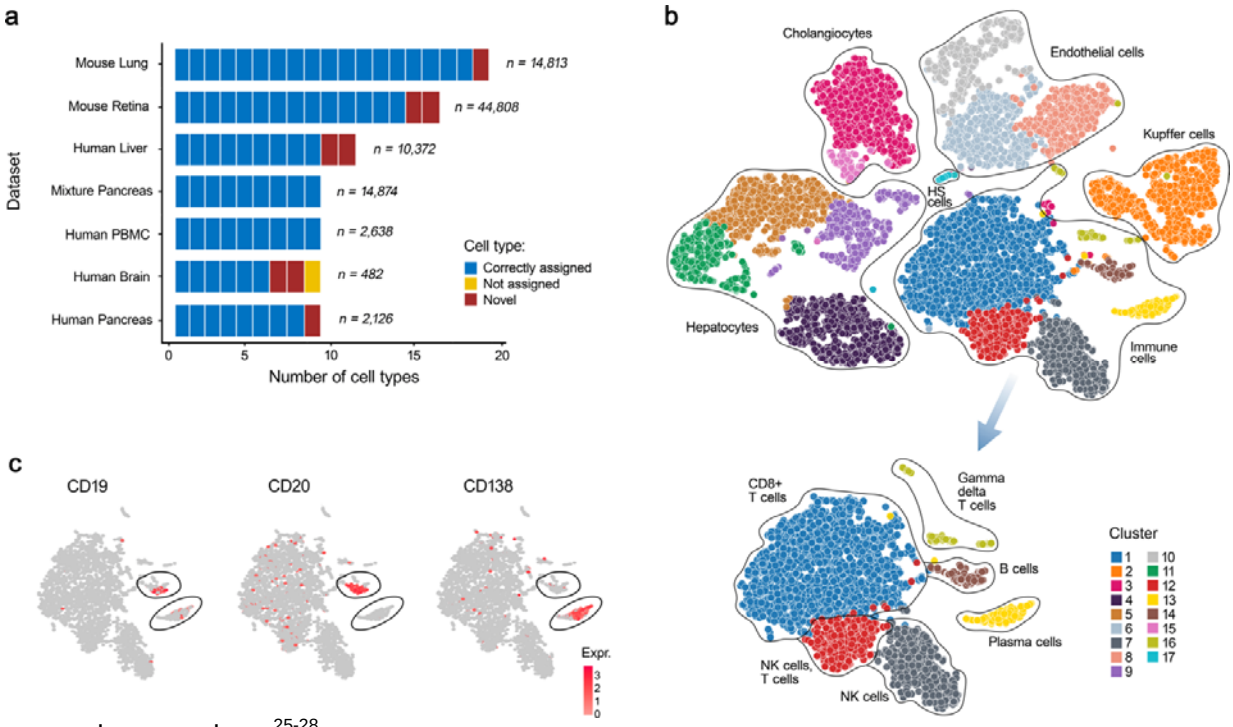


Figure 2. ScType identifies cluster- and cell-type-specific markers across multiple datasets. (**a**) The overall performance of ScType across 7 human and mouse scRNA-seq datasets. ScType automatically assigned cell-types similar to the original studies in human datasets, and it also correctly reannotated 5 cell types in the brain, liver and pancreas tissues. Similarly, ScType did not only enable automated identification all the cell types in mouse lung and retina, but it also correctly reassigned three cell types in those datasets. ScType labelled only single cluster (fetal cells) as

unknown cell type in the human brain dataset. **(b**) A more detailed example of cell subtype identification by ScType in the liver atlas dataset, where it automatically labelled the same cell-types as assigned in the original manuscript.[8] (**c**) ScType improved the cell type annotation of specific subclasses of liver atlas dataset. Two different cell clusters that were annotated as B-cells in the original study were clearly segregated into B-cell and plasma (B) cell types, as plasma cells do not express common B-cell markers, such as CD19 and CD20, but instead express CD138.

ScType improves automated cell type assignment over existing methods

We further compared how the selection and number of the marker genes affects the accuracy of cell-type annotation with both ScType and commonly-used differential expression analysis-based approach. For the comparison, we identified an increasing number of differentially expressed (DE) genes (standard approach), as well as cell type- and cluster-specific markers (ScType approach) for each detected cell cluster (see Methods), both based on the same unsupervised clustering of the 7 scRNA-seq datasets. In the standard approach, the identified top DE genes were matched to the CellMarker database[6], while in the data-driven ScType approach, the top specific marker genes were matched to the ScType database to assign the cell-types. In 4 out of 7 datasets, the top-4 marker genes from ScType approach were enough for correct cell-type assignment of all 38 clusters, while top-10 markers (default value) allowed correct cell-type annotation of 81 out of 82 cell clusters in the 7 datasets (Fig 3a). Sub-optimal performance was observed only in the human brain dataset due to mislabelling of fetal brain cells.

In contrast, when using the top DE genes for each cluster, the median percentage of correctly-assigned clusters remained below 50% until using the maximum of 50 genes (Fig. 3b). The standard approach also showed more variability between the datasets. For instance, the performance of the DE-based approach in the human pancreas dataset[17] was higher (74% accuracy), as these cell-types selectively expresses specific genes (e.g. insulin by β-cells), which have been extensively-studied as markers for decades, and these genes are well-captured by the CellMarker database[6]. Since the standard approach performed best in the human pancreatic cells, we compared the cluster-specific expression markers selected using both approaches in the human pancreas dataset. The expression heatmaps of the top ScType markers offer a more informative visual cluster separation of the distinct cell populations, such as α and β-cells clusters (Supplementary Fig. 1a), when compared to expression heatmaps based on the DE marker genes (Supplementary Fig. 1b).

Finally, we compared the unsupervised ScType against a supervised cell-type classification method, CellAssign[29]. Instead of assigning cell-types to clusters defined by unsupervised clustering, CellAssign uses a probabilistic Bayesian model to determine the likelihood of each cell belonging to a cell-type defined by user-provided set of marker genes. In the human pancreas dataset, ScType correctly annotated all the 9 cell types (Fig. 3c), while CellAssign, together with

tissue-specific marker genes from CellMarker database,[6] was able to correctly annotate only gamma, acinar and ductal cells (Fig. 3d). This is because the performance of any prior knowledge-based cell classifier depends strongly on the selected set of markers[30], which are challenging to define for certain tissue and cell types. The running time of ScType was only 5 seconds, after clustering and DE detection steps that took ca. 6 minutes on a standard desktop machine, whereas running of CellAssign took 85 minutes. The CellAssign failed to run in the other 6 datasets due to the large number of known markers in the CellMarker database and large number of cells in the scRNA-seq datasets (see Methods).

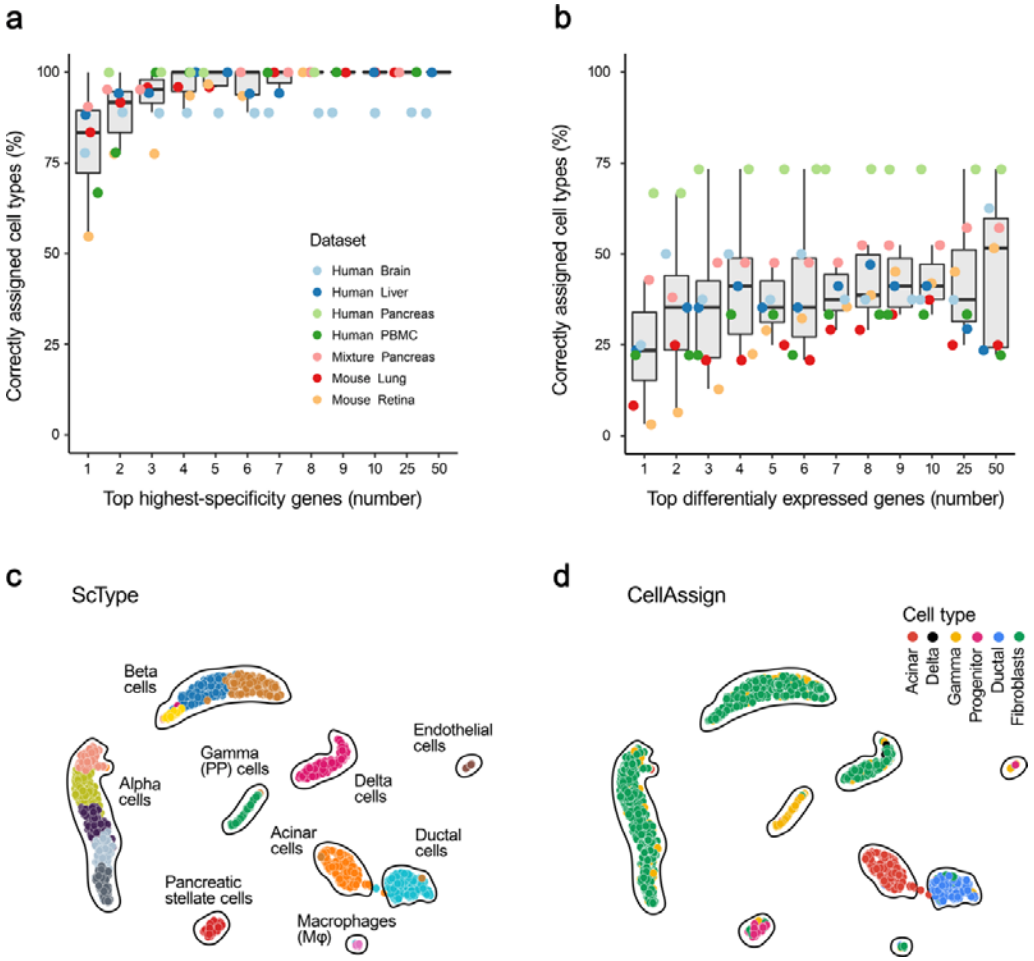


Figure 3: Comparison of ScType with standard analysis and CellAssign to automatically label cell-types. (**a**) Boxplots show the percentage of correctly-assigned cell-types to detected clusters using an increasing number of top cell type- and cluster-specific marker genes extracted by ScType. (**b**) Boxplots show the percentage of correctly-assigned cell-types to clusters using an increasing number of most differentially-expressed genes (standard approach). (**c**) Unsupervised ScType automatically identified the human pancreas cell subpopulations[17] using the top-10 cell type- and cluster-specific marker genes for each cell cluster (default option). (**d**) Supervised CellAssign method[29] that uses a prior knowledge of known cell markers was able to correctly and

specifically annotate only the groups of gamma, acinar and ductal cells in the human pancreas dataset[17].

Discussion

We presented ScType, an automated cell type- and cluster-specific marker gene selection method which allows accurate single-cell-type annotations based solely on the given scRNA-seq data. To promote its wide application, either as a stand-alone tool or together with other popular single-cell data analysis tools (e.g., Seurat, MAST, PAGODA), we have implemented ScType both as an interactive web-platform (http://sctype.fimm.fi), and as an open-source R implementation (https://sctype.fimm.fi/source_code.php). We anticipate the method will accelerate unbiased phenotypic profiling of cells when applied either to large-scale single-cell sequencing projects or smaller-scale molecular and functional profiling of patient-derived samples. For example, the integrative marker information in the ScType database may enable the identification of rare molecular cell subtypes that have distinct combinations of markers, suggesting specific molecular functions in the body.

The existing computational methods for automatic identification of cell types can be broadly categorized into two groups: (1) supervised methods that require annotated training datasets labelled with correct cell populations to train the classifiers (e.g. CaSTLe[31] and ACTINN[32] that annotate cell types based on pre-defined reference set of cell without the need of cell marker input), and (2) a prior knowledge-based methods that require either a marker gene set or a pre-trained classifier for selected cell populations (e.g. Garnett[33] that utilizes first marker genes to identify representative cell types and then trains a regression model to classify the remaining cells to one of the cell types). Although a recent comparison showed that supervised methods outperformed prior knowledge-based methods[30], supervised methods may have severe limitations when annotating rare populations of cells due to lack of reference data to train the machine learning algorithms. Furthermore, supervised methods are notoriously time-consuming to train as well as error prone, as technical artifacts in the training data affect their prediction ability for new scRNA-seq data.

Similarly, the prior knowledge-based cell classification approaches have certain limitations. For instance, their performance heavily depends on the available gene lists provided as markers for each cell type, typically obtained from manual literature search or matching to marker databases that are still sub-optimal both in coverage and specificity. Ideally, one would like to use an appropriate number of specific markers to achieve a maximally accurate and precise cell-type classification. However, most existing methods utilize a limited number of markers, thereby potentially masking the identification of a subpopulation of cells that do not express the selected marker genes. Furthermore, the use of inconsistent cell type markers across experiments and

laboratories may compromise the reproducibility of the findings[30]. These caveats become even more pronounced as the number of cell types and samples increases, thus preventing fast and reproducible annotations. It has been therefore argued that prior information does improve the automated cell-type identifications.[30]

ScType implements a number of improvements compared to the existing cell-annotation tools. Our unsupervised approach outperformed CellAssign, a marker gene-based probabilistic cell-type assignment method, which was recently shown to enable accurate annotation of multiple cell types[29]. Another group of supervised methods, such as CaSTLe[31], ACTINN[32], SingleR[34] and CHETAH[35], utilize reference bulk or single-cell transcriptomic data for cell type predictions, and therefore require comprehensive, manually-annotated and high-quality reference datasets; furthermore, these tools do not allow identification of novel cell-type marker genes. In contrast, ScType requires neither reference scRNA-seq datasets nor manual selection of marker genes; instead, all the background information for established or *de novo* markers comes from the novel ScType database that is to date the most comprehensive database of specific markers for human and mouse cells.

In comparison with many other computational methods that require manual interference,[29,33] ScType takes a data-driven and a marker-independent approach, and it annotates the cell-types at once in a single-cell experiment in a totally unsupervised manner. The only input needed for the ScType tool is the raw sequencing data file, although uploading of pre-processed scRNA-seq data is also an option. This saves considerable time and costs in the scRNA-seq analysis, especially when searching for cell-types in a tissue that involves large variety of cell-types with similar transcriptomic profiles (e.g. bone marrow samples from mixed lineage leukemia subjects). Additionally, ScType score allows identification of novel marker genes with high specificity for either known or new cell types. For example, the algorithm enables one to flag those genes that show high cluster-specific expression in a particular cell type but which have not yet been reported in the cell marker databases.

Using 7 scRNA-seq datasets from human and mouse tissues, we demonstrated that ScType provides scalable and accurate identification of cell-clusters and is compatible with data formats from various sequencing techniques (e.g. Drop-seq and Smart-seq). These benchmarking results against the existing cell annotation approaches indicated that ScType is a widely-applicable to various biomedical problems. Further, the comprehensive ScType database may lead to the development of new and improved cell-type detection methods, as well as accelerate the implementation of single-cell pipelines for translational applications, such as monitoring of therapy resistant cancer cell sub-populations, which require fast and automated analyses. As more scRNA-seq datasets from various tissue types become available from the Human Cell Atlas

and other projects, the accuracy and coverage of the ScType tool and database is expected to increase accordingly.

Authors contributions

AI, AKG and TA conceived and planned the study. AI developed the method, implemented the ScType web-tool and collected and analyzed the data. AI compiled the ScType database with the help of AKG. AI prepared the figures for manuscript with the help of TA and AKG. All the authors wrote the manuscript and approved its final version.

Acknowledgements

The authors thank Dr. Pirkko M Mattila, Dr. Jenni Lahtela and Bhiswa Ghimire for their valuable suggestions to improve the web-tool, and Olle Hansson for the cluster server machine to host the web-tool and the database. This work was supported by the Academy of Finland (grants 292611, 295504, 310507, 326238), European Union's Horizon 2020 Research and Innovation Programme (ERA PerMed JAKSTAT-TARGET), the Cancer Society of Finland (TA) and the Sigrid Jusélius Foundation (TA).

References

1. Pellin, D. et al. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat Commun 10, 2395, doi:10.1038/s41467-019-10291-0 (2019).
2. Cui, Y. et al. Single-cell transcriptome analysis maps the developmental track of the human heart. Cell Rep 26, 1934-1950 e1935, doi:10.1016/j.celrep.2019.01.079 (2019).
3. Maestre-Batlle, D. et al. Novel flow cytometry approach to identify bronchial epithelial cells from healthy human airways. Sci Rep 7, 42214, doi:10.1038/srep42214 (2017).
4. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat Methods 9, 743-748, doi:10.1038/nmeth.2069 (2012).
5. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods 3, 793-796, doi:10.1038/nmeth929 (2006).
6. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 47, D721-D728, doi:10.1093/nar/gky900 (2019).
7. Franzen, O., Gan, L. M. & Bjorkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) 2019, doi:10.1093/database/baz046 (2019).
8. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature 572, 199-204, doi:10.1038/s41586-019-1373-2 (2019).
9. van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell 176, 1265-1281 e1224, doi:10.1016/j.cell.2019.01.031 (2019).
10. Regev, A. et al. The Human Cell Atlas. Elife 6, doi:10.7554/eLife.27041 (2017).
11. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20, 273-282, doi:10.1038/s41576-018-0088-9 (2019).
12. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 15, e8746, doi:10.15252/msb.20188746 (2019).
13. Macosko, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

14. Kim, D. S., Matsuda, T. & Cepko, C. L. A core paired-type and POU homeodomain-containing transcription factor program drives retinal bipolar cell gene expression. J Neurosci 28, 7748-7764, doi:10.1523/JNEUROSCI.0397-08.2008 (2008).
15. Kim, J. W. et al. Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals. Dev Cell 37, 520-532, doi:10.1016/j.devcel.2016.05.023 (2016).
16. Cherry, T. J. et al. Development and diversification of retinal amacrine interneurons at single cell resolution. Proc Natl Acad Sci USA. 2009 Jun 9;106(23):9495-500. doi: 10.1073/pnas.0903264106 (2009).
17. Muraro, M. J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 3, 385-394 e383, doi:10.1016/j.cels.2016.09.002 (2016).
18. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 8, 14049, doi:10.1038/ncomms14049 (2017).
19. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A 112, 7285-7290, doi:10.1073/pnas.1507125112 (2015).
20. Angelidis, I. et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. Nat Commun 10, 963, doi:10.1038/s41467-019-08831-9 (2019).
21. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019).
22. Herzog, E. et al. Expression of vesicular glutamate transporters, VGLUT1 and VGLUT2, in cholinergic spinal motoneurons. Eur J Neurosci 20, 1752-1760, doi:10.1111/j.1460-9568.2004.03628.x (2004).
23. Dong, H. et al. Excessive expression of acetylcholinesterase impairs glutamatergic synaptogenesis in hippocampal neurons. J Neurosci 24, 8950-8960, doi:10.1523/JNEUROSCI.2106-04.2004 (2004).
24. Tellier, J. & Nutt, S. L. Standing out from the crowd: How to identify plasma cells. European Journal of Immunology 47, 1276-1279, doi:10.1002/eji.201747168 (2017).
25. Faust, D. Modulation of C1q mRNA Expression and Secretion by Interleukin-1, Interleukin-6, and Interferon-g in Resident and Stimulated Murine Peritoneal Macrophages. Immunobiology 206, 368-376, doi:10.1078/0171-2985-00187 (2002).
26. Zhao, Y. et al. The immunological function of CD52 and its targeting in organ transplantation. Inflamm Res 66, 571-578, doi:10.1007/s00011-017-1032-8 (2017).
27. Haasken, S. et al. Macrophage scavenger receptor 1 (Msr1, SR-A) influences B cell autoimmunity by regulating soluble autoantigen concentration. J Immunol 191, 1055-1062, doi:10.4049/jimmunol.1201680 (2013).
28. Westerterp, M. et al. Apolipoprotein CI aggravates atherosclerosis development in ApoE-knockout mice despite mediating cholesterol efflux from macrophages. Atherosclerosis 195, e9-16, doi:10.1016/j.atherosclerosis.2007.01.015 (2007).
29. Zhang, A. W. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods 16, 1007-1015, doi:10.1038/s41592-019-0529-1 (2019).
30. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194, doi:10.1186/s13059-019-1795-z (2019).
31. 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, doi:10.1371/journal.pone.0205499 (2018).

32. Ma, F. & Pellegrini, M. ACTINN: Automated Identification of Cell Types in Single Cell RNA Sequencing. Bioinformatics, doi:10.1093/bioinformatics/btz592 (2019).
33. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. Nat Methods 16, 983-986, doi:10.1038/s41592-019-0535-3 (2019).
34. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 20, 163-172, doi:10.1038/s41590-018-0276-y (2019).
35. 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, doi:10.1093/nar/gkz543 (2019).
36. Tenenbaum, J.B. et al. A global geometric framework for nonlinear dimensionality reduction. Science. 22;290(5500), 2319-23 (2000).
37. Coifman R.R. et. al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci. 24;102(21), 7426-31 (2005).
38. Tang, J. et al. Visualizing Large-scale and High-dimensional Data. Proc. 25th Int. Conf. World Wide Web 287–297 (2016). doi:10.1145/2872427.2883041
39. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16, 278, doi:10.1186/s13059-015-0844-5 (2015).
40. Abadi M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, arXiv:1603.04467

ONLINE METHODS

ScType database construction

We have built the largest database to date of human and mouse cell-specific markers by integrating the information available in the CellMarker database (http://biocc.hrbmu.edu.cn/CellMarker/) and PanglaoDB (https://panglaodb.se). CellMarker and PanglaoDB are currently the two largest available databases for cell type markers. However, these two databases differ in the number of tissues, cell types and marker numbers, as well as in the way the markers have been assigned to each cell type. In case of CellMarker database, 13 605 cell markers for 467 cell types in 158 human tissues/sub-tissues and 9148 cell makers for 389 cell types in 81 mouse tissues/sub-tissues were manually collected and curated from more than 100 000 published papers[6]. In the PanglaoDB, 6631 gene markers mapping to 155 cell types have been identified by differential expression analysis in particular cell types using single cell data and a community-based crowdsourcing approach for curation of gene expression markers[7]. Therefore, we firstly converted the non-uniform gene IDs to approved gene symbols within and between the databases. Next, we removed the low evidence marker genes from CellMarker database (genes having only one reference for being a certain cell type marker), and genes that appeared in less than 5 clusters of specific cell type from PanglaoDB. Additionally, we excluded genes showing no expression across all the datasets in PanglaoDB. Ultimately, we unified the cell and tissue naming from the two databases and excluded tissues comprising less than 5 cell types. Fifteen novel cell types with corresponding marker genes were added by manual curation of multiple papers to the current version of the compiled ScType database (https://sctype.fimm.fi/database.php), as relatively few brain and eye tissue cell types were provided in the first version of the database. For instance, the current version of ScType database comprises 3980 cell markers for 194 cell types in 17 human tissues and 4212 cell markers for 194 cell types in 17 mouse tissues. Cell-type specificity was calculated separately for every marker gene across the cell types, hence providing

a quantitative measure of how frequently the marker identifies the cell type uniquely within the tissue using the cell-type specificity score (Eq. 1).

Publicly available datasets

In order to benchmark the ScType against the other approaches, we utilized 7 scRNA-seq datasets from public domain and re-analysed these data using ScType. Five datasets were downloaded from Gene Express Omnibus (GEO): Human Liver (GSE124395), Human Brain (GSE67835), Human Pancreas (GSE85241), Mouse Lung (GSE63269) and Mouse Retina (GSE63473). Human PBMC dataset was downloaded from the 10x Genomics Dataset Repository (https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz). The Human Pancreas Mixture dataset is a combination of datasets downloaded using the following accession numbers: GSE81076, GSE85241, GSE86469, E-MTAB-5061, and GSE84133.

The scType workflow options

ScType provides a complete pipeline for single-cell RNA-seq data analysis and cell-type annotation. We utilized Seurat v3.1.0 for data processing and normalization. For clustering analysis, the default option is Louvain clustering based on a shared nearest neighbour graph (using FindClusters function with the resolution parameter set to 0.8 and 20 principal components given as input), which was used to generate the current results; however, also SC3, DBSCAN, GiniClust and k-means clustering options are available in ScType. The clusters are visualized using either principal components analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), Isomap[36], Diffusion Map[37], largeVis[38] or by means of expression heatmaps. The differential expression analysis between each identified cluster (reference cluster) versus all the other detected clusters was performed using MAST[39] (default option), or using non-parametric Wilcoxon Rank Sum test (non-parametric test). In the current results, we used MAST to identify an increasing number of top differentially-expressed genes with the highest positive fold-change (FDR<0.05 or maximum of 50 genes). Based on these identified genes, separate scores for cell type-specificity (s_{type}) and cells cluster-specificity (s_{clust}) were calculated as shown in Eqs. (1) and (2). The top-10 genes among the selected genes with the highest marker-specificity scores (s^i) (Eq. 3) were used to identify cell types by matching with the marker-cell type information in ScType database (or using user-provided custom cell type gene sets as an alternative option). The cumulative sum of marker-specificity scores of all the genes supporting a cell-type is used as the final score (so-called ScType score) to tag a label to the cluster. Sctype score is calculated for each cell-type within the tissue, and the cell label with the highest ScType score is assigned to the cluster. In addition to cell type assignments, the ScType web-portal (http://sctype.fimm.fi) allows users to view the metadata based on which the assignment was made, view the markers that are enriched in each specific cluster, and plot the cumulative gene-specificity for different cell types as bar graphs. For the integrated, multi scRNA-seq dataset analysis, ScType uses FindIntegrationAnchors and IntegrateData functions from Seurat v3.1.0 that were shown to enable an effective identification of anchor correspondences across multiple single-cell datasets[21].

Identification of specific markers

For the automated cell type annotation, ScType utilizes the top cluster- and cell type-specific marker genes. ScType calculates cell-type-specificity score for a particular gene g_i and a user-specified tissue type ($s_{type}^{g_i}$) based on information in the ScType database as follows:

$$s_{type}^{g_i} = \frac{\text{number of celltypes with } g_i \text{ as marker in the specified tissue}}{\text{total number of cell types in the specified tissue}} \quad (1)$$

Cell-type-specificity score (s_{type}) of equals to 0 when the gene is a known marker for all the cell types, and 1 when the gene is a known marker only for a specific cell type within the specified tissue. Cell-type-specificity score allows selection of genes that are specific to certain cell type within a tissue.

The cluster-specificity score ($s_{cluster}^{g_i}$) for gene g_i and particular cluster is calculated as follows:

$$s_{cluster}^{g_i} = \frac{\text{number of cells in a selected cluster where } g_i \text{ is expressed}}{\text{total number of cells in the dataset where } g_i \text{ is expressed}} \quad (2)$$

Only those cells with a normalized expression above 5th percentile of overall gene expression in the dataset (excluding cells with zero expression) are counted as cells expressing the particular gene. Cluster-specificity score (s_{clust}) equals to 0 when the gene is expressed in any or all cell clusters except for the given cluster, and 1 when the gene is expressed only in the cells of the particular cell cluster. s_{clust} allows the selection of genes that are highly expressed uniquely in a given cell cluster.

Ultimately, ScType calculates cell type- and cluster-specificity score (s^i) for each selected gene using the geometric mean of the cell-type-specificity (s_{type}) and cluster-specificity ($s_{cluster}$) scores:

$$s^i = \sqrt{s_{type}^{g_i} \, s_{cluster}^{g_i}} \quad (3)$$

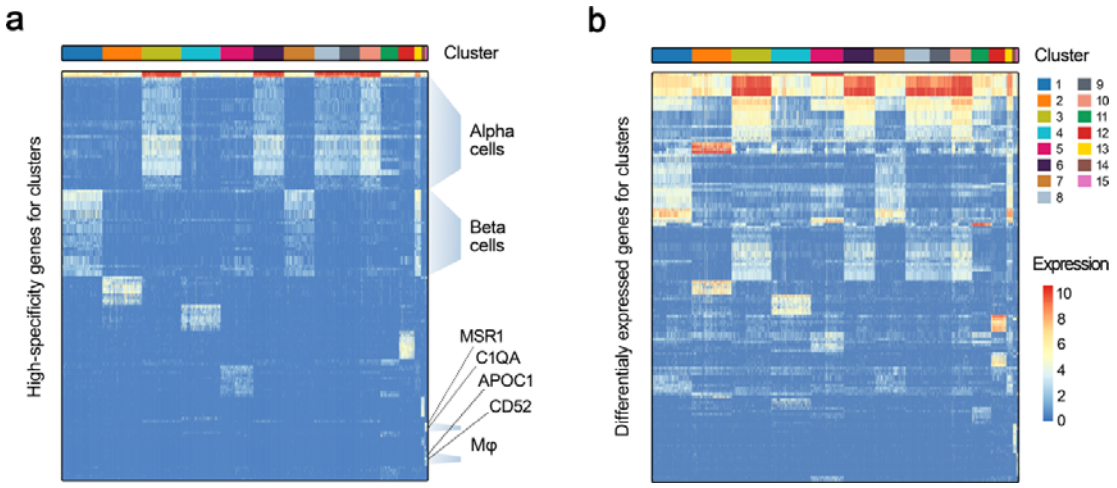
Comparison with CellAssign

We compared the accuracy, speed and requirement of hardware resources of ScType against CellAssign[29] using the 7 scRNA-seq datasets used in the study. We used default parameters to run CellAssign in the human pancreas dataset, and note that CellAssign, implemented using Google's TensorFlow[40], failed to execute when applied to other 6 datasets considered in the study, where it reported an "Exhausted error" when allocating tensor (on Intel Core i5-8250U 3.4-GHz machine with 96 GB RAM, 16GB of RAM plus 80GB of swap space), due to the large number of known cell type markers in the CellMarker database[6] and a large number of cells in the scRNA-seq datasets.

Code and data availability

The R source-code of the ScType algorithm is freely available at https://sctype.fimm.fi/source_code.php to allow reproduction of the results and its further comparison against or integration with other algorithms. ScType is also freely available as an interactive web-tool at http://sctype.fimm.fi. The ScType database is freely available at https://sctype.fimm.fi/database.php.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Expression heatmaps of markers from ScType and standard approach in human pancreas dataset.[17] **(a)** Expression heatmap of the top-10 cell type- and cluster-specific marker genes for the detected clusters extracted with ScType. C1QA, APOC1, CD52 are the top markers that were used to annotate small macrophages (Mφ) cell population (see Fig. 3c). (**b**) Expression heatmap of the top-10 differentially-expressed genes for the detected clusters.