HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2022 ; 27: 199–210.

CAMML: Multi-Label Immune Cell-Typing and Stemness Analysis for Single-Cell RNA-sequencing

Courtney Schiebout, **H. Robert Frost**

Biomedical Data Science, Dartmouth College, Lebanon, NH 03766, USA

Abstract

Inferring the cell types in single-cell RNA-sequencing (scRNA-seq) data is of particular importance for understanding the potential cellular mechanisms and phenotypes occurring in complex tissues, such as the tumor-immune microenvironment (TME). The sparsity and noise of scRNA-seq data, combined with the fact that immune cell types often occur on a continuum, make cell typing of TME scRNA-seq data a significant challenge. Several single-label cell typing methods have been put forth to address the limitations of noise and sparsity, but accounting for the often overlapped spectrum of cell types in the immune TME remains an obstacle. To address this, we developed a new scRNA-seq cell-typing method, Cell-typing using variance Adjusted Mahalanobis distances with Multi-Labeling (CAMML). CAMML leverages cell type-specific weighted gene sets to score every cell in a dataset for every potential cell type. This allows cells to be labelled either by their highest scoring cell type as a single label classification or based on a score cut-off to give multi-label classification. For single-label cell typing, CAMML performance is comparable to existing cell typing methods, SingleR and Garnett. For scenarios where cells may exhibit features of multiple cell types (e.g., undifferentiated cells), the multi-label classification supported by CAMML offers important benefits relative to the current state-of-the-art methods. By integrating data across studies, omics platforms, and species, CAMML serves as a robust and adaptable method for overcoming the challenges of scRNA-seq analysis.

Keywords

single-cell RNA sequencing; gene set testing; cell typing; stemness

1. Introduction

The development of single cell methods, in particular single cell transcriptomics, has dramatically changed the landscape of omics research in the past decade.[1,2] In particular, single cell analysis enables researchers to characterize the full range of cell types/states present in a tissue.[3,4] Understanding what cell types are present, and in what proportions, is critical in many single-cell RNA-sequencing (scRNA-seq) experiments. In cancer models, distinguishing immune cell types and their proportions is key to understanding mechanisms

Courtney.T.Schiebout.GR@dartmouth.edu .

of immune evasion and tumor growth, and thus how to inhibit them.[5,6] Additionally, cell typing is vital to the characterization of healthy tissues in the development of single-cell atlases of organs or biological systems.[7] While manual curation of cell types by an expert is possible in these examples, often the volume of cells and complexity of tissues undergoing scRNA-seq analysis makes the use of an automated cell-typing tool highly preferable, if not essential.[2,3,8,9]

Cell-typing of scRNA-seq data is often done using cell clustering results, with investigators selecting the most likely cell type for each cluster based on differentially expressed genes.[8–10] However, this approach fails to account for cluster heterogeneity, i.e, a cluster containing several similar cell types, and is sensitive to the selection of clustering algorithm and algorithm parameters.[11,12] To address these issues, several cluster-independent methods for cell-typing (e.g., SingleR and Garnett) have been recently developed that utilize expression profile correlation or machine learning models to predict cell types.[13–15] Although these methods are easy to use and less biased than manual cluster-based approaches, they have an important limitation. Specifically, the cell-typing methods currently available are all single-label, i.e., they provide just a single cell type classification for each cell.[13–15] Cell types, especially immune cell types, often occur on a spectrum.[16,17] If two cell types are both highly plausible classifications for a cell, methods that restrict the classification to only one cell type may do so incorrectly while also eliminating potentially useful information regarding cell phenotypes from alternative classifications. A related challenge occurs when no cell types fit a given cell with high confidence. In this case, current single-label methods may classify the cell with its most likely cell type anyway, providing potentially false information that may hinder scRNA-seq analyses.

Current cell-typing methods are further limited by the perception that cell subtypes are mutually exclusive. Cell subtypes are often identified by the distinct proteins present on the cell surface and/or by cell location within an organism.[18,19] However, the gene expression profile for one cell subtype is often not highly distinct from the profiles of related subtypes.[20] This presents two challenges for cell-subtyping of scRNA-seq data. Since the actual gene expression of cells in different subtypes may not be easily differentiable, accurately identifying a single label for these subtypes using transcriptomic data is extremely challenging.[16,17] Additionally, identifying these subtypes may not be informative for cancer response if the associated transcriptomes almost completely overlap. Rather, understanding the phenotypic nature of single cells and how they contribute to model function and dysfunction may be more informative.

To ameliorate these issues, we developed a new cell-typing method for scRNA-seq data, Cell-typing using variance Adjusted Mahalanobis distances with Multi-Labeling (CAMML). CAMML uses customizable weighted gene sets for each cell type of interest and calculates Variance-adjusted Mahalanobis (VAM)[21] scores for these cell type gene sets for each cell in a scRNA-seq dataset. These scores allow for cells to not only be classified by their most likely cell type, but also enable the use of multi-label classification. Because the scores generated by CAMML can be transformed into valid p-values under the null hypothesis of uncorrelated technical noise, CAMML can be tailored to only give classifications with high specificity according to multiple hypothesis testing, allowing for greater confidence in

cell-typing results. The generation of scores for each potential cell type also enables the evaluation of cell stemness/differentiation, i.e, cells with comparable scores among several cell types can be considered more stem-like and cells that score high for a single cell type can be considered more highly differentiated. In the remainder of the paper, we present our CAMML method and compare its performance to other cell-typing methods on scRNA-seq data with known cell type labels. An R package implementing the CAMML method is in development and will be released on CRAN, but for now it is available at https://github.com/schiebout/CAMML.

2. Methods

2.1. Gene Set Development

2.1.1. Public Expression Data—To test CAMML's cell-typing method, several single-cell RNA-seq datasets with annotated cell types were accessed. The datasets included scRNA-seq data for 10 fluorescence-activated cell sorted (FACS)[18] immune cell populations from Zheng,[22] available on the 10X Genomics website (https://www.10xgenomics.com/resources/datasets), and immune cell scRNA-seq data from a manually curated melanoma dataset from Gene Expression Omnibus (GEO), available at accession GSE72056.[23,24] These scRNA-seq datasets were processed using Seurat v.4.01 in R v.4.0.2.[25] Cells with over 5% of reads belonging to mitochondrial genes were removed, as were genes present in fewer than 100 cells and cells with fewer than 500 genes. Log normalization was then applied and Seurat's nearest neighbor algorithm was used over 30 dimensions to perform unsupervised clustering with a resolution of 0.25.[25] Data was visualized using Uniform Manifold Approximation and Projection (UMAP) on the top 30 principal components.[25,26]

2.1.2. Gene Set Optimization—To build gene sets that distinguish cell types, differential gene expression analysis was performed on reference expression data from the R package celldex,[13] which includes cell typelabeled human and murine bulk gene expression data. Specifically, the Human Primary Cell Atlas (HPCA) was used to generate cell type gene sets by performing one vs. all differential gene expression analysis using the exact test in edgeR v.3.32.1.[13,27,28] To define cell type gene set membership, a differential expression cutoff was used to determine which up-regulated genes would be included in a given cell type gene set. For our analysis, this cutoff was based on the log fold-change in expression of the gene between the target cell type and all other cell types. The cutoff used for each analyzed dataset that spanned basal cell types was set to 5 for high stringency. In the case of analyzing cells across subtypes, a lower threshold of 3 was used as expression across subtypes does not differ enough to result in genes with high log fold-changes. These differential expression-based gene sets were refined using cell type gene sets from the C8 collection of the Molecular Signatures Database (MSigDB).[29] Most C8 cell type profiles were obtained from the Hay bone marrow gene sets;[30] for cell types not included in the Hay sets, gene sets were obtained from the C8 heart gene sets.[31] The intersection of these MSigDB gene sets and the differential expression gene sets was used to identify genes that were consistently associated with each cell type across study conditions, and this was used as the final gene set for each cell type. The median cell type gene set size was 25 genes.

2.2. Cell-Typing

2.2.1. Variance-Adjusted Mahalanobis Distance—CAMML generates cell type scores using a modified version of the Variance-Adjusted Mahalanobis (VAM) method,[21] which we developed to support cell-level gene set scoring of noisy and sparse scRNA-seq data. For this application, VAM is applied to two input matrices: **X**: $n \times p$ matrix that holds the normalized scRNA-seq counts for p genes in n cells, and **A**: $m \times p$ matrix that represents the annotation of the p genes to m gene sets representing distinct cell types. To capture both gene set membership and gene weights, element $a_{i,j}$ of **A** will be 0 if gene j is not included in the set for cell type i or the positive weight of gene j for cell type i. Using **X** and **A**, the modified VAM method computes an $n \times m$ matrix **S** as follows:

1. **Compute modified Mahalanobis distances**: Let **M** be an $n \times m$ matrix of squared values of a modified version of the Mahalanobis multivariate distance measure.[32] Each column k of **M**, which holds the cell-specific squared distances for cell type k, is calculated as $\mathbf{M}[,k] = \text{diag}\,\mathbf{M}[,k] = diag\left(\mathbf{X}_k^T\left(\mathbf{I}_g\hat{\sigma}^2_{k,\ \text{tech}}\right)^{-1}\mathbf{X}_k\right)$, where g is the gene set size for cell type k, $\mathbf{X}_k$ is a $n \times g$ matrix containing the g columns of **X** corresponding to the members of the set for cell type k, $\mathbf{I}_g$ is a $g \times g$ identity matrix, and $\hat{\sigma}^2_{k,\ \text{tech}}$ holds the ratio of the technical variance of the g genes in set k to cell type-specific weights for the g genes.
2. **Compute modified Mahalanobis distances on permuted X**: To capture the distribution of the squared modified Mahalanobis distances under the H_0 that the normalized expression values in **X** are uncorrelated with only technical variance, let $\mathbf{X_p}$ represent the row-permuted version **X** and let $\mathbf{M_p}$ be the $n \times m$ matrix that holds the squared modified Mahalanobis distances computed on $\mathbf{X_p}$.
3. **Fit gamma distribution to each column of $\mathbf{M_p}$**: A separate gamma distribution is fit via maximum likelihood to the non-zero elements in each column of $\mathbf{M_p}$.
4. **Use gamma cumulative distribution function (CDF) to compute cell-specific scores**: The cell-specific gene set scores are set to the CDF value for each element of **M**.

The use of a CDF to generate the elements of **S** has several important benefits: 1) it transforms the squared modified Mahalanobis distances for gene sets of different sizes into a common scale, which is important if values in **S** are used together in statistical models, e.g., as regression predictors, 2) it generates a statistic that is bound between 0 and 1 and is robust to very large expression values, i.e., the CDF converges quickly to 1 as the squared distances increase, and 3) valid p-values can be generated by subtracting the elements of **S** from 1.

2.2.2. Basal Cell-Typing—Half of the cell types in Zheng (2017) and all of the cell types in GSE72056 can be considered basal immune cell types, such as macrophages, T cells, and B cells.[22,23] To train CAMML on the most broadly applicable cell types, the model was first developed to identify these basal cell types. In both datasets, after the development of the necessary cell type gene sets, weighted VAM scores were calculated following the procedure detailed above for each cell across each cell type. With these

scores, cells were classified in several ways. First, the highest scoring cell type for each cell was used as a single-label designation, allowing for a comparative evaluation against SingleR and Garnett.[13,15] Second, the two highest scoring cell types for each cell were used to evaluate if the ”ground truth” cell type was captured among the highest scoring cell types, if not the highest scoring cell type. Multi-labeled cells were then visualized using UMAP dimensionality reduction to determine if trends of continuous expression could be detected.[26]

2.2.3. Cell-Subtyping—Following basal cell typing, the Zheng (2017) data was further analyzed to determine cell subtypes.[22] The Zheng dataset has sequencing information for 6 different types of T cells (CD4+, CD8+, naive CD4+, naive CD8+, memory, and regulatory T cells), enabling more specific characterization.[22] New gene sets for each T cell subtype were developed using the same differential expression approach employed for basal cell types on a subset of the training data containing just T cells. The subtype gene sets were based on a differential expression analysis restricted to T cells in order to prioritize genes that help distinguish one T cell subtype from other subtypes rather than genes that separate T cells from non-T cells. The MSigDB C8 collection does not currently have gene sets for all of these T cell subtypes and was thus excluded from the gene set building in this case. VAM scores were then computed using the weighted T cell subtype gene sets for all cells previously classified as basal T cells. The accuracy of the CAMML method for T cell subtyping was compared to existing methods, SingleR and Garnett, based on single-label classification using the highest scoring T cell subtype and multi-label classification using the same approach described in the basal cell-typing methods.[13,15]

2.3. Stemness

2.3.1. Public Time Series scRNA-seq Data—To gauge if multi-label cell typing could be leveraged to evaluate stemness, public time series scRNA-seq datasets were utilized. The first dataset utilized was accessed via GEO, at accession number GSE118068.[24,33] This study performed scRNA-seq on embryonic mouse cerebellums at 10, 12, 14, 16, and 18 days of gestation, as well as on newborn mouse cerebellums at 0, 5, 7, and 14 days post birth.[33] The second dataset (GEO accession number GSE107122) contained scRNA-seq data generated on embryonic mouse brains at embryonic stages 11.5, 13.5, 15.5, and 17.5 days.[24,34] The same scRNA-seq computational processing steps previously described for the cell-typing datasets were used for these time series datasets.

2.3.2. Entropy—As a proxy for stemness, an entropy-based measure was calculated for each of the cells in the aforementioned datasets. This was accomplished using a similar process to that employed for the multi-label cell typing evaluation. First, gene sets were created for each relevant mouse cell type using the differential expression method outlined above applied to the mouse RNA-seq reference dataset in celldex.[13,35] MSigDB gene sets were not used in this case as they are specific to human tissues. VAM scores were then calculated for each of the weighted cell type gene sets on each of the cells in the public murine scRNA-seq datasets. After VAM scores were computed, the cell type entropy of each cell, in this case defined by a modified Shannon Diversity Index (mSDI) (outlined in Equation 1), was calculated. The Shannon Diversity Index was modified in order to account

for the strength of the VAM scores, so that cells with only one nonzero cell type score did not have identical mSDIs if one had a much lower VAM cell type score than another. These entropy scores were then evaluated using a linear model in R.

$$H = -\sum_{i=1}^{R} p_i \ln p_i$$
$$p_i = \frac{VAM_i + \epsilon}{\sum VAM} \quad (1)$$

3. Results and Discussion

3.1. Cell-typing Performance

To evaluate the accuracy of CAMML relative to existing methods, we performed cell-typing using single-label CAMML, multi-label CAMML for the top two cells types, Garnett,[15] and SingleR[13] on the aforementioned public scRNA-seq datasets from 10X Genomics and GEO.[22,23] For each method, the processing and filtering of the public scRNA-seq data being tested and the cell types available for potential classification were identical, although Garnett does include an additional classification option of ”unknown” that CAMML and SingleR do not support.[13,15] Furthermore, both the Garnett and SingleR classifiers were trained on data provided in their documentation as each method required a specific and unique structure for training data.[13,15] In the case of SingleR, the training data was also based on HPCA; however, for Garnett, a manually curated training dataset was used.[13,15] The motivation for evaluating CAMML both as a single-label and multi-label classifier was to test how it performed compared to other methods when restricted to a single label and to assess how multi-label classification may further inform cell-typing. Although not pictured, AUCell was also used for basal cell-typing to compare multi-label CAMML to another method capable of multi-label cell typing.[36] The gene sets built for CAMML were fed into AUCell and cell-type classification was determined based on the default method for classification built into AUCell.[36] In each case, multi-label CAMML performed comparably or better than AUCell, confirming the utility of CAMML as a multi-label classifier and the robustness of the gene sets we built.

As displayed in Figure 1a, single- and multi-label CAMML and SingleR performed with similar accuracy when tested on the Zheng dataset.[13,22] Of note, the CD34+ cells in this dataset reported a FACS sorting purity of about 50%, which likely explains why all methods struggle with this cell type.[22] Figure 1b similarly illustrates that all methods perform relatively consistently on the GEO melanoma dataset,[23] with one notable exception. CAMML has a reduced accuracy for T cells in this dataset, particularly single-label CAMML. Further evaluation of the method performance in this cell type category indicated that CAMML detects a crossover of expression between T cell and NK cell gene sets, which is discussed in more detail in the multi-labelling section. Figure 1c displays the accuracy of cell-subtyping performed on the T cell subtypes available in the Zheng dataset.[22] In most cases, single and multi-label CAMML performed better than SingleR.[13] Of note, Garnett[15] was not used in this example because it does not currently have T cell subtype training

data available. Both CAMML methods and SingleR fail to achieve accuracy comparable to that observed in the basal cell-typing, supporting findings in previous literature that T cell subtypes often occur on a continuum of expression and may not be easily distinguishable from transcriptomic data.[16,17] In both methods, cell types of similar origins were often all classified into one subtype, resulting in high accuracy of one subtype and relatively low accuracies in others. In all subtypes, the use of multi-labelling improves the accuracy of CAMML, supporting our hypothesis that multi-labelling provides more information and context when carrying out scRNA-seq analysis.

3.2. Multi-Labelling

In order to evaluate the reduced accuracy observed for T cells in the GSE72056 data[23] and to assess the benefits of a multi-label classifier in immune TME cell-typing, UMAP visualization was used to further study the similarity in gene expression between T cells and NK cells.[26] Cells labelled by the original authors as T cells or NK cells were isolated, and Figure 2 shows these cells projected onto the first two UMAP dimensions. Figure 2a shows the VAM scores for T cells, NK cells, and an overlay of both, as well as a fourth panel illustrating the color key.[21,25] Figure 2b shows what the cells were actually identified as in the original study.[23]

Cells identified as NK cells in the original study do not score highly for T cells, which is further supported by their near-perfect accuracy across single- and multi-label CAMML and SingleR as visualized in Figure 1b.[13,23] However, many of the cells identified as T cells have high VAM scores for both T cells and NK cells. Given that cytotoxic T cells and NK cells often have overlapping gene expression and immune action, this overlap is not surprising.[20,37] However, being able to detect this crossover through multi-label cell-typing provides context that would otherwise be lost. It appears from this data that many T cells are expressing genes that align with the cytotoxic profile of NK cells, potentially indicating that there is notable T cell activation occurring in this TME. In comparative studies of cancer therapies in the immune compartment of the TME, the use of multi-label cell-typing may aid in understanding how immune cells are transitioning and altering their phenotypes in response to cancer.

3.3. Stemness Analysis

Following CAMML's performance with cell-typing, the potential for a second assessment measure emerged. We hypothesized that undifferentiated or stem-like cells may be responsible for cases where CAMML did not score strongly for any cell type. To evaluate the validity of this hypothesis, two public time series scRNA-seq datasets were evaluated for change in mSDI (Equation 1) as a proxy for measuring cell differentiation over time. Figure 3 shows the results of this study, with about 60,000 cells across 9 time points from GEO dataset GSE118068[33] in Figure 3a and 11,000 cells across 4 time points from GEO dataset GSE107122[34] in Figure 3b.

Despite a consistent median mSDI, both datasets had significant decreases over time ($p <$ .001) when modeled by univariate linear regression. As visualized in Figure 3a and 3b, the first time point has many fewer cells with low mSDI scores than other time points, leading

us to theorize that although the bulk of cells still have high mSDIs, certain cells are dropping in mSDI over time as they become differentiated. Furthermore, given that both datasets come from mouse scRNA-seq datasets that focus entirely on or start with embryonic cells, it is not surprising that most cells are still not completely differentiated throughout the time series data as mice brains are still developing.[33,34,38,39] Interestingly, in Figure 3a, the linear process of differentiation also appears to slightly restart during the postnatal period, with postnatal day 0 having fewer low mSDI cells than embryonic day 18, and the low mSDI cells reemerging in later postnatal time points. In both cases, there is a significant trend in decreasing mSDI scores over time, indicating the potential for CAMML to capture meaningful patterns in stemness.

4. Conclusion

Single-cell RNA-sequencing has enabled characterization of tissues that was previously unattainable with bulk transcriptomics methods. However, the challenge of identifying the cell types present in a tissue and their proportions has been an ongoing challenge, particularly as a result of the sparsity and noise present in scRNA-seq data. Several approaches have been developed to ameliorate this issue. Manual clustering methods are a common way to assign cell types; however, the potential bias of performing cell-typing manually and the assumption that all cells within a computationally-defined cluster are the same cell type have limited the effectiveness of these methods. To overcome this, more complex computational methods, such as SingleR[13] and Garnett[15] have been developed to take an automated, unbiased approach to scRNA-seq cell-typing. While these are promising methods for identifying single cell types, in cases where cell types are continuous rather than discrete, these methods fall short of fully characterizing the complexities of the tissue being studied. This challenge is especially relevant for the tumor-immune microenvironment, where expression of cell types can occur on a spectrum and cells can alter their phenotype in response to cancer and immune signaling.[16,17]

In order to overcome the issue of continuous cell types while maintaining the effectiveness of single-label cell-typing supported by existing methods, we developed the Cell-typing using variance Adjusted Mahalanobis distances with Multi-Labeling (CAMML) method, a novel method that uses weighted cell type gene sets and a statistical technique optimized for sparse and noisy scRNA-seq data to score individual cells for multiple cell types.[21] Using cell type gene sets generated on existing public data sources, including MSigDB[29] and celldex,[13] CAMML performs comparably or better than existing methods in terms of single-label classification accuracy. Furthermore, while we used fixed log fold-change thresholds of 5 for basal cells types and 3 for subtypes in our development of the gene sets for these analyses, users of CAMML with prior expertise on their cell types of interest could customize their cutoff in order to maximize their accuracy. In future, we also plan to develop a downregulated cell-type option, where genes with low or no expression in a certain cell type can be considered alongside those that are upregulated in order to improve specificity. By incorporating a multi-label option, CAMML provides additional context for cell phenotypes not available with existing cell typing methods, as illustrated in Figure 2, where the inclusion of a second label highlighted a group of T cells with a strong similarity to NK cells, suggesting a cytotoxic phenotype. This multi-label can be used to forego the

traditional option of only assigning one cell type to any single cell. However, it can also be used to inform phenotype even when a single label is chosen. Given that the VAM[21] uses a CDF to score cells, a p-value is available for each cell type in each cell, allowing for informed cell classification. The versatility of CAMML is further promoted by its potential use in stemness analysis, whereby a modified SDI can be used to assess a cell's level of differentiation. The incorporation of multiple datasets into the training and testing of CAMML, combined with its diverse analysis capabilities, make it a robust and promising method for analyzing scRNA-seq data, particularly in the immune compartment of the TME.

Acknowledgments

We thank Drs. Christensen, Huang, and Salas for their helpful discussion and support.

Reviewer Contributors

We thank the reviewers, Almut Lütge, Fan Zhang, and Van Truong, for their contributions.

References

1. Aldridge S. & Teichmann SA Single cell transcriptomics comes of age. Nature Communications 11, 4307 (2020).
2. Svensson V, Vento-Tormo R. & Teichmann SA Exponential scaling of single-cell RNA-seqin the past decade. Nature Protocols 13, 599–604 (2018). [PubMed: 29494575]
3. Diaz-Mejia JJ et al. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. F1000Research 8, 296 (2019). URL https://f1000research.com/articles/8-296/v3.
4. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell 58, 610–620 (2015). URL https://www.cell.com/molecular-cell/abstract/S1097-2765(15)00261-0. Publisher: Elsevier. [PubMed: 26000846]
5. Hanahan D. & Coussens LM Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell 21, 309–322 (2012). [PubMed: 22439926]
6. Gajewski TF, Schreiber H. & Fu Y-X Innate and adaptive immune cells in the tumor microenvironment. Nature immunology 14, 1014–1022 (2013). URL https://pubmed.ncbi.nlm.nih.gov/24048123https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4118725/. [PubMed: 24048123]
7. Wagner J. et al. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell 177, 1330–1345.e18 (2019). URL http://www.sciencedirect.com/science/article/pii/S0092867419302673. [PubMed: 30982598]
8. Zeisel A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (New York, N.Y.) 347, 1138–1142 (2015).
9. Pollen AA et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature biotechnology 32, 1053–1058 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191988/.
10. Kiselev VY, Andrews TS & Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics 20, 273–282 (2019). URL 10.1038/s41576-018-0088-9.
11. Duo A, Robinson MD & Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research 7, 1141 (2020). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134335/.
12. Freytag S, Tian L, Lönnstedt I, Ng M. & Bahlo M. Comparison of clustering tools in R formedium-sized 10x Genomics single-cell RNA-sequencing data. F1000Research 7, 1297 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124389/. [PubMed: 30228881]

13. Aran D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology 20, 163–172 (2019). URL 10.1038/s41590-018-0276-y. [PubMed: 30643263]
14. de Kanter JK, Lijnzaad P, Candelli T, Margaritis T. & Holstege FCP CHETAH: aselective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic acids research 47, e95–e95 (2019). URL https://pubmed.ncbi.nlm.nih.gov/31226206https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895264/. Publisher: Oxford University Press. [PubMed: 31226206]
15. Pliner HA, Shendure J. & Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nature Methods 16, 983–986 (2019). URL 10.1038/s41592-019-0535-3. [PubMed: 31501545]
16. Li H. et al. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell 176, 775–789.e18 (2019). URL 10.1016/j.cell.2018.11.043. Publisher: Elsevier. [PubMed: 30595452]
17. Zhang Q. et al. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. Cell 179, 829–845.e20 (2019). URL 10.1016/j.cell.2019.10.003. Publisher: Elsevier. [PubMed: 31675496]
18. Bonner WA, Hulett HR, Sweet RG & Herzenberg LA Fluorescence activated cellsorting. The Review of Scientific Instruments 43, 404–409 (1972). [PubMed: 5013444]
19. Woodland DL & Kohlmeier JE Migration, maintenance and recall of memory T cells in peripheral tissues. Nature Reviews Immunology 9, 153–161 (2009). URL https://www.nature.com/articles/nri2496. Bandiera abtest: a Cg type: Nature Research Journals Number: 3 Primary a type: Reviews Publisher: Nature Publishing Group.
20. Judge SJ et al. Differences in NK and Memory CD8 T cell responses to antigen-nonspecific stimulation by interleukin-15. The Journal of Immunology 204, 148.17–148.17 (2020). URL https://www.jimmunol.org/content/204/1_Supplement/148.17. Publisher: American Association of Immunologists Section: Cellular Mechanisms of Innate Immunity: NK Cells, ILCs, and Other Non-Macrophages.
21. Frost HR Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene set scoring. Nucleic Acids Research 48, e94–e94 (2020). URL 10.1093/nar/gkaa582. [PubMed: 32633778]
22. Zheng GXY et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049 (2017). URL 10.1038/ncomms14049.
23. Tirosh I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (New York, N.Y.) 352, 189–196 (2016).
24. Edgar R, Domrachev M. & Lash AE Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30, 207–210 (2002). [PubMed: 11752295]
25. Satija R, Farrell JA, Gennert D, Schier AF & Regev A. Spatial reconstruction of single-cell gene expression data. Nature Biotechnology 33, 495–502 (2015). URL 10.1038/nbt.3192.
26. McInnes L, Healy J. & Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [cs, stat] (2020). URL http://arxiv.org/abs/1802.03426. ArXiv: 1802.03426.
27. Mabbott NA, Baillie JK, Brown H, Freeman TC & Hume DA An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics 14, 632 (2013). URL 10.1186/1471-2164-14-632. [PubMed: 24053356]
28. Robinson MD, McCarthy DJ & Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010). URL 10.1093/bioinformatics/btp616. [PubMed: 19910308]
29. Liberzon A. et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740 (2011). URL 10.1093/bioinformatics/btr260. [PubMed: 21546393]
30. Hay SB, Ferchen K, Chetal K, Grimes HL & Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. Experimental Hematology 68, 51–61 (2018). [PubMed: 30243574]

31. Cui Y. et al. Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. Cell Reports 26, 1934–1950.e5 (2019). [PubMed: 30759401]
32. Mahalanobis PC On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta) 2, 49–55 (1936).
33. Vladoiu MC et al. Childhood Cerebellar Tumors Mirror Conserved Fetal Transcriptional Programs. Nature 572, 67–73 (2019). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6675628/. [PubMed: 31043743]
34. Yuzwa SA et al. Developmental Emergence of Adult Neural Stem Cells as Revealed by Single-Cell Transcriptional Profiling. Cell Reports 21, 3970–3986 (2017). URL https://www.sciencedirect.com/science/article/pii/S2211124717318132. [PubMed: 29281841]
35. Benayoun BA et al. Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. Genome Research 29, 697–709 (2019). [PubMed: 30858345]
36. Aibar S. et al. SCENIC: single-cell regulatory network inference and clustering. Nature Methods 14, 1083–1086 (2017). URL http://www.nature.com/articles/nmeth.4463. [PubMed: 28991892]
37. Narni-Mancinelli E, Vivier E. & Kerdiles YM The ‘T-cell-ness’ of NK cells: unexpected similarities between NK cells and T cells. International Immunology 23, 427–431 (2011). URL 10.1093/intimm/dxr035. [PubMed: 21665959]
38. Semple BD, Blomgren K, Gimlin K, Ferriero DM & Noble-Haeusslein LJ Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species. Progress in neurobiology 0, 1–16 (2013). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737272/.
39. Chen VS et al. Histology Atlas of the Developing Prenatal and Postnatal Mouse Central Nervous System, with Emphasis on Prenatal Days E7.5 to E18.5. Toxicologic Pathology 45, 705–744 (2017). URL 10.1177/0192623317728134. Publisher: SAGE Publications Inc. [PubMed: 28891434]

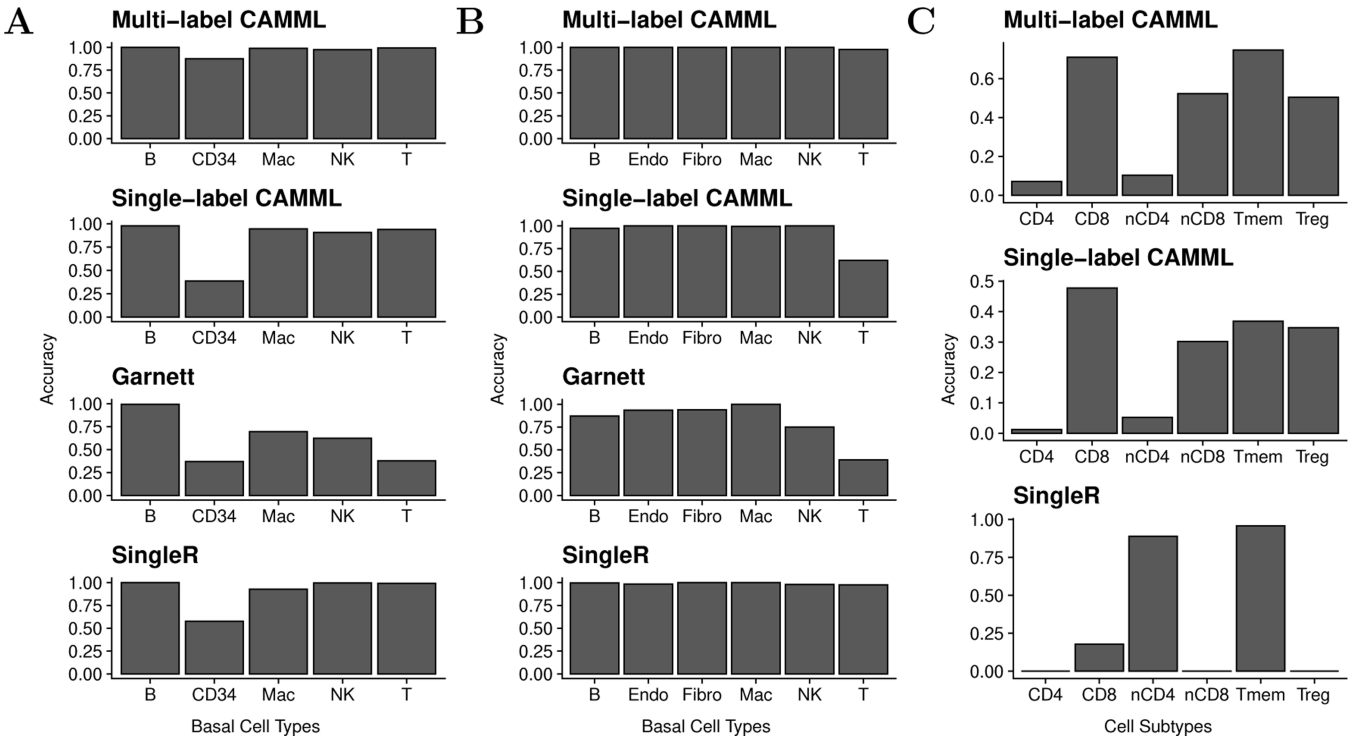

Fig. 1.

Classification accuracy of CAMML on the top 2 scoring cell types and the top scoring cell type, Garnett, and SingleR for (**A**) the FACS-sorted 10X Genomics dataset on B cells, CD34+ hematopoietic stem cells, macrophages, NK cells, and T cells,[22] (**B**) the melanoma public scRNA-seq dataset[23] on B cells, endothelial cells, fibroblasts, macrophages, NK cells, and T cells, and (**C**) on the T cell subsets identified in the FACS-sorted 10X dataset: CD4+ T cells, CD8+ T cells, naive CD4+ T cells, naive CD8+ T cells, memory T cells, and regulatory T cells.

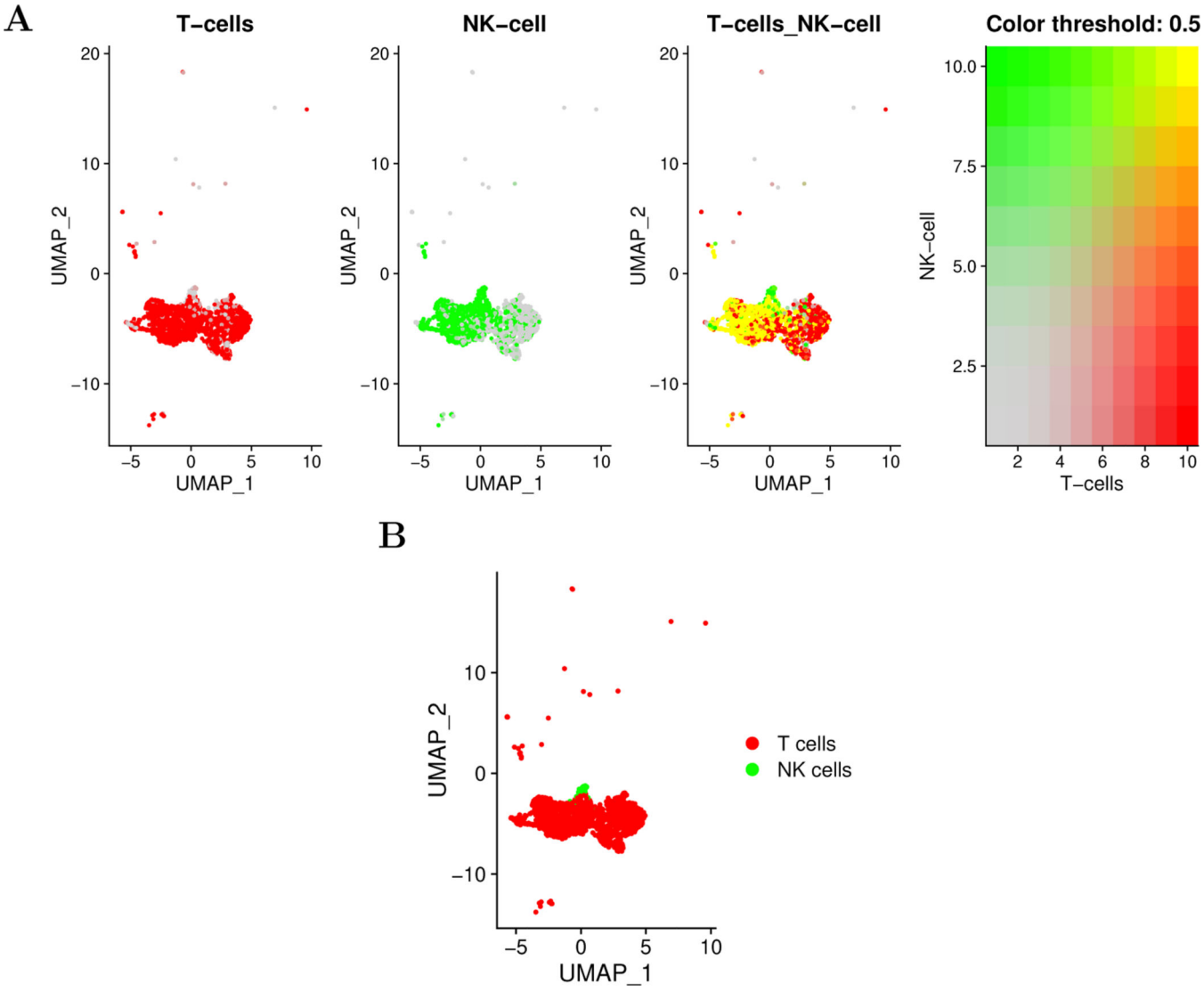

Fig. 2.

Visualizations of cells classified as T cells or NK cells in the GEO GSE72056 melanoma dataset[23] on the first two UMAP dimensions, (**A**) colored by weighted VAM scores for T cells, NK cells, and both, with a color key, and (**B**) colored by cell type assignments from the original study.[23]

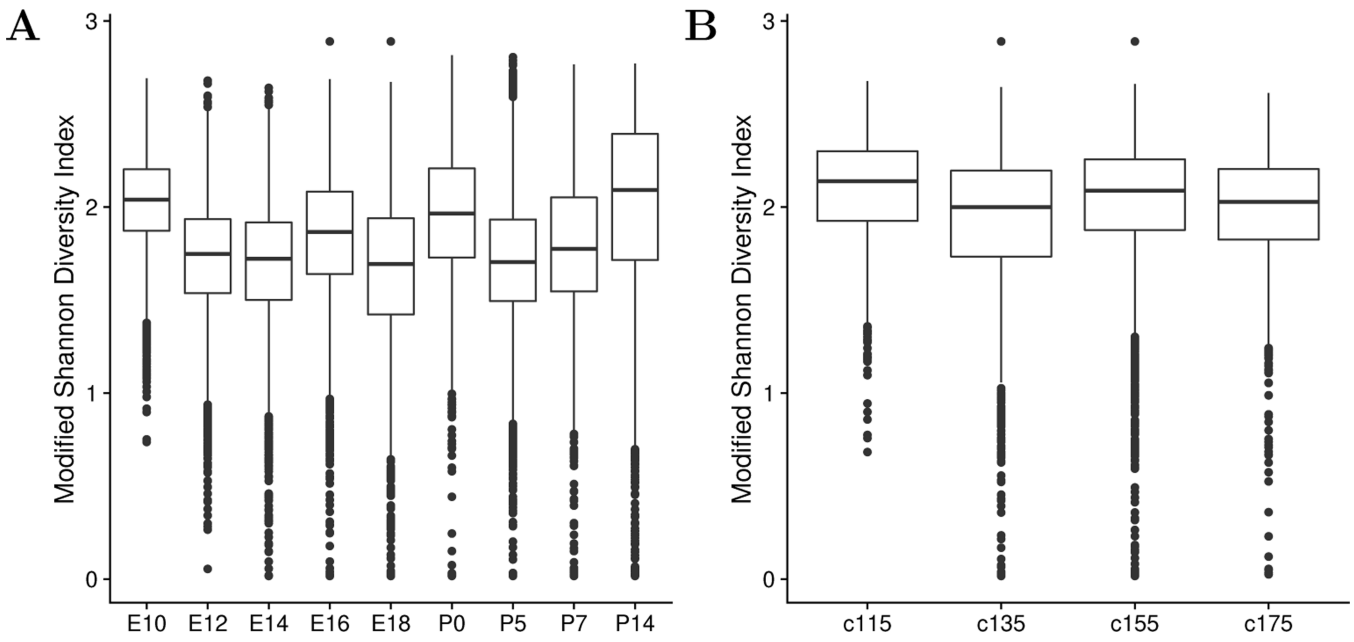

Fig. 3.

Boxplots of mSDI (Eq. 1) across time in **A** GSE118068 data, with mouse cerebellum scRNA-seq data from embryonic timepoints of 10, 12, 14, 16, and 18 days of gestation and postnatal timepoints of 0, 5, 7, and 14 days old,[33] and in **B** GSE107122 data of embryonic mouse brains at embryonic stages 11.5, 13.5, 15.5, and 17.5.[34]