Title

Cell type matching in single-cell RNA-sequencing data using FR-Match

Authors: Yun Zhang[1], Brian Aevermann[1], Rohan Gala[2], Richard H. Scheuermann[1,3,4]

Affiliations: [1]J. Craig Venter Institute, La Jolla, CA, USA; [2]Allen Institute for Brain Science, Seattle, WA, USA; [3]Department of Pathology, University of California San Diego, La Jolla, CA, USA; [4]Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA, USA

Abstract

Reference cell type atlases powered by single cell transcriptomic profiling technologies have become available to study cellular diversity at a granular level. We present FR-Match for matching query datasets to reference atlases with robust and accurate performance for identifying common and novel cell types and suboptimally clustered cell types in the query data. FR-Match shows excellent performance for cross-platform, cross-sample type, and cross-tissue region cell type matching.

Main

Single cell transcriptomic profiling has emerged as a powerful tool to characterize the cellular heterogeneity in complex biological systems. Large collaborative consortia, including the Human Cell Atlas [1] and NIH BRIAN Initiative [2, 3], have adopted the unbiased single-cell/nucleus RNA-sequencing (sc/snRNA-seq) technologies to generate reference cell type atlases at single cell resolution across many organs and species at an unprecedented level of granularity; a series of recent publications have reported 128 transcriptomically-distinct cell types in human primary motor cortex (M1) [4], 116 cell types in mouse primary motor cortex (MOp) [5], and 75 cell types in human middle temporal gyrus (MTG) neocortex [6]. The Allen Institute for Brain Science has made these comprehensive datasets available to serve as reference cell type atlases (https://portal.brain-map.org/atlases-and-data/rnaseq).

An important role for these reference datasets is to support the matching of new query data to the reference cell types using recently-developed computational methods. Azimuth is a web application for reference-based single-cell analysis following the Seurat pipeline [7]. Online iMNF is an extension of the Liger pipeline for single-cell multi-omics integration using iterative online learning [8, 9]. ScArches is a deep learning strategy for mapping query datasets on top of a reference by single-cell architectural surgery [10]. The mathematical foundation of these methods are linear algebra techniques (canonical correlation analysis (CCA) for Seurat and non-negative matrix factorization (NMF) for Liger) that effectively decompose the structure of large data matrices for integrative analysis. While these methods are great tools for single-cell data integration, producing integrated UMAP visualization for both query and reference datasets with minimal batch effects, cell type matching is a more pragmatic use case that requires not only integrating the query cells into the reference, but also being able to make a clear distinction between common and novel cell types existing in the query dataset and the studied conditions.

Previously, we reported a computational pipeline for downstream cell type analysis of scRNA-seq data combining NS-Forest, a random forest machine learning algorithm for the identification of minimum sets of marker genes for given cell types [11, 12], and FR-Match, a minimum spanning tree-based statistical testing approach for cell type matching of query and reference datasets, and demonstrated its performance in matching cell types from overlapping brain regions [13]. We introduced the concept of

cell type "barcodes" [12, 13] using NS-Forest marker genes to visualize and characterize the distinction between different cell types [12]. The NS-Forest marker genes also serve as a feature selection approach for FR-Match that essentially matches the query and reference cell types based on the cell type barcode gene expression patterns [13]. Here, we report recent enhancements to FR-Match, including a normalization step and a cell-to-cluster matching scheme, and show that the cell type barcodes provide evidence and explainability for the matching results. The enhanced FR-Match was found to effectively match cell types between platforms (10X and SMART-seq; scRNA-seq and smFISH), sample types (whole cells and nuclei), and tissue regions (human M1 and MTG) and provided evidence for suboptimal partitioning in the clustering step.

We designed a normalization procedure based on the marker gene expression patterns observed in the cell type barcode plots, to dampen technical artifacts observed in different scRNA-seq platforms. We observed that barcode plots from the SMART-seq platform (Figure 1A(i)) and the 10X platform (Figure 1A(iv)) showed similar marker gene expression specificity, but different expression distributions (Supplemental Figure 1) and variable non-specific background expression. To address these technical artifacts for cell type matching, min-max rescaling is applied to each gene independently for both SMART-seq and 10X data, to globally align the data in the range of [0, 1]. The SMART-seq platform showed better sensitivity for low expression genes than the 10X experiment, but also showed more background noise. To reduce the background noise while preserving the expression signals, the normalization step for the SMART-seq data uses a per-barcode per-gene summary statistic (mean or median) to weight the barcode pattern by multiplying the summary statistic to the gene expression submatrix of that cluster (see Methods). In the case of median, the result if that for genes expressed in a minority of cells (median = 0), expression in these cells is set to zero. Finally, the weighted barcode is again rescaled to [0, 1] for matching. The above procedure effectively aligned the cross-platform barcode patterns (Figure 1A(ii)(iii)), producing similar signal and noise levels.

To augment the original cluster-to-cluster matching in FR-Match, a cell-to-cluster extension was added based on an iterative procedure that allows each cell in the query cluster to be assigned a summary p-value, quantifying the confidence of matching to a reference cluster (see Methods). This extension is available as a stand-alone function "FRmatch_cell2cluster()" in the "FRmatch" R package (https://github.com/JCVenterInstitute/FRmatch). A cosine distance option was also added for robust matching between experiments with global data variabilities.

Using these pipelines and extensions, the cross-platform matching approach was validated using Allen human M1 snRNA-seq data generated using the *10X Chromium v3* protocol [4] as the reference and an M1 snRNA-seq dataset from another Allen study on multiple human cortical regions using the *SMART-seq v4* protocol [6] as a query. Although the raw counts of the query and the reference data showed very different data distributions (Supplementary Figure 1) the FR-Match matching results produced almost all one-to-one matching at the subclass level for all query cells (Figure 1B), with the exception of the agglomerated IT query type. Due to the grouping (under-partitioning) of the layer-non-specific IT cells in the query, the majority of these cells were matched to one of the different layer-specific IT reference types. These results verify that the normalization step for aligning SMART-seq and 10X data is effective and the extended FR-Match is robust to perform cross-platform cell type matching.

We also applied this FR-Match pipeline to assess cross-sample type matching using an *snRNA-seq* dataset from the Allen mouse MOp [5] as reference cell types and an *scRNA-seq* dataset from the MOp subset of a cell type taxonomy of the entire adult mouse isocortex and hippocampus [14] as the query. Since both datasets were generated using the 10X protocol, we only applied the min-max scaling in the

normalization step. For subclass types, most of the query types were one-to-one matched to a reference type (Figure 1C). The highlighted box shows that query SMC-Peri cells were matched to either the SMC or Peri types in the reference, with ~50:50 split. A one-to-many match may indicate that the query cluster is under-partitioned [13]. An examination of the cell type barcode plots for these query and reference cell types (Figure 1D) showed two distinct patterns in the query barcode, each corresponding to one of the two reference barcodes supporting the under-partitioning hypothesis. Thus, the FR-Match cell type matching pipeline, together with the cell type barcode, showed excellent matching of single nucleus and single cell clusters and provided solid evidence of suboptimal partitioning based on marker gene expression in the matching results.

We benchmarked the FR-Match pipeline with the Azimuth and Online iNMF approaches for identification of suboptimal parititoning using human M1 (Supplementary Figure 2A-B) and mouse MOp (Supplementary Figure 2C-D) datasets. All cells were matched, but these integration methods were not able to split the under-partitioned clusters. In the mouse MOp case, Azimuth matched all query SMC-Peri cells to the reference Peri subclass with a few mismatched to the reference VLMC subclass (Supplementary Figure 2C). The Online iNMF produced joint clustering of the integrated data instead of explicitly reporting the cell-to-cell mapping. All the query SMC-Peri cells and the reference SMC and Peri cells were grouped in the same cluster from the joint clustering (Supplementary Figure 2D).

For the above two use cases, we also matched the most granular cell types and benchmarked in comparison with Azimuth. In the mouse MOp use case, the FR-Match results formed a clean diagonal alignment of cell types and assigned unmatched cells as “unassigned” in the bottom row (Figure 2A); in some cases with no matches to the reference clusters suggesting the presence of novel cell types in the query data. The Azimuth results also showed the majority of the matching cells along the diagonal, but with many more suboptimal matches scattered off-diagonal and no indication of novel unassigned cell types in the query data (Figure 2B). Similar results for human M1 can be found in Supplementary Figure 3.

Another important matching challenge is to match cell types across different tissues or anatomic regions within a tissue. The cluster-to-cluster version of FR-Match [13] was used to match cell clusters from the SMART-seq platform between the human M1 [4] and MTG [6] brain regions and the bi-directional (M1 as query to MTG as reference, and vice versa) matching results are shown in Figure 2C. Most of the GABAergic neuron and all of the glial cell types were nicely matched across these two cortical brain regions, but none of the glutamatergic neuron types were matched. This suggests that the inhibitory neuron and glial cell types are conserved across brain regions, whereas the excitatory neurons are cortical region specific. We also examined the cell type barcode plots for pairs of matched cell types showing highly similar expression patterns of the matched types using reciprocal marker genes, even though the best marker gene sets selected for each brain regions may be different when defined based on the cell types present in the dataset (e.g., Supplementary Figure 4).

Finally, we report the application of FR-Match for matching spatial transcriptomics data generated by single molecular fluorescence in situ hybridization (smFISH) [15] to a SMART-seq scRNA-seq dataset as the reference, both from mouse primary visual cortex (VISp) [16]. *De novo* clustering of the smFISH data was used to obtain broadly-defined cell type clusters. The FR-Match cell-to-cluster pipeline was used to assign a reference cell type to each spatial cell in the initial clusters. The FR-Match results successfully recapitulated the clear laminar distributions of excitatory neurons, corresponding to their assigned cell types (Figure 2D-E). In contrast, the inhibitory neurons were scattered across all layers, with the *Vip* type located more densely in upper layers and the *Sst* and the *Pvalb* types located more

densely in deeper layers. The FR-Match cell type assignment for the spatially sequenced cells reflected their expected laminar patterns.

In summary, we extended our cell type matching pipeline to perform both cell-to-cluster and cluster-to-cluster matching. The added normalization step and cosine distance option allow FR-Match to perform robust and accurate cell type matching across platforms (SMART-seq with 10X), sample types (single-cell with single-nucleus), brain regions (M1 with MTG), and data modalities (spatial transcriptomics with scRNA-seq). Compared with other methods, FR-Match can effectively detect non-optimally partitioned clusters from the previous clustering step, and uniquely identify potential novel cell types as "unassigned" cells. The cell type barcodes can be useful for investigators to interpret the underlying transcriptomic drivers of the FR-Match results, assisting future research directions.

References (up to 20 references)

1. Regev, A., et al., *The Human Cell Atlas.* Elife, 2017. **6**.
2. *The impact of the NIH BRAIN Initiative.* Nat Methods, 2018. **15**(11): p. 839.
3. Insel, T.R., S.C. Landis, and F.S. Collins, *The NIH brain initiative.* Science, 2013. **340**(6133): p. 687-688.
4. Bakken, T.E., et al., *Comparative cellular analysis of motor cortex in human, marmoset and mouse.* Nature, 2021. **598**(7879): p. 111-119.
5. Yao, Z., et al., *A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex.* Nature, 2021. **598**(7879): p. 103-110.
6. Hodge, R.D., et al., *Conserved cell types with divergent features in human versus mouse cortex.* Nature, 2019. **573**(7772): p. 61-68.
7. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
8. Gao, C., et al., *Iterative single-cell multi-omic integration using online learning.* Nature Biotechnology, 2021: p. 1-8.
9. Welch, J.D., et al., *Single-cell multi-omic integration compares and contrasts features of brain cell identity.* Cell, 2019. **177**(7): p. 1873-1887. e17.
10. Lotfollahi, M., et al., *Mapping single-cell data to reference atlases by transfer learning.* Nature Biotechnology, 2021: p. 1-10.
11. Aevermann, B.D., et al., *Cell type discovery using single-cell transcriptomics: implications for ontological representation.* Human molecular genetics, 2018. **27**(R1): p. R40-R47.
12. Aevermann, B., et al., *A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing.* Genome research, 2021. **31**(10): p. 1767-1780.
13. Zhang, Y., et al., *FR-Match: robust matching of cell type clusters from single cell RNA sequencing data using the Friedman–Rafsky non-parametric test.* Briefings in Bioinformatics, 2020.
14. Yao, Z., et al., *A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation.* Cell, 2021. **184**(12): p. 3222-3241. e26.
15. Lein, E., L.E. Borm, and S. Linnarsson, *The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing.* Science, 2017. **358**(6359): p. 64-69.
16. Tasic, B., et al., *Shared and distinct transcriptomic cell types across neocortical areas.* Nature, 2018. **563**(7729): p. 72-78.
17. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of statistical mechanics: theory and experiment, 2008. **2008**(10): p. P10008.

18. Traag, V.A., L. Waltman, and N.J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Sci Rep, 2019. **9**(1): p. 5233.
19. Friedman, J.H. and L.C. Rafsky, *Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests.* The Annals of Statistics, 1979: p. 697-717.
20. Ding, J., et al., *Systematic comparison of single-cell and single-nucleus RNA-sequencing methods.* Nature biotechnology, 2020. **38**(6): p. 737-746.

Methods

FR-Match cell-to-cluster matching algorithm

As originally conceived, FR-Match is a cluster-to-cluster matching algorithm that utilizes a graphical model and minimum spanning trees to determine the data distributional equivalence between two cell type clusters derived from single cell or single nucleus RNA sequencing (scRNA-seq) data in multivariate space [13]. The required input data for FR-Match are cell-by-gene expression matrices and cell cluster membership labels for both query and reference data. The output of the original FR-Match is a map between the query cluster and the reference cluster labels, thus assigning known reference cell types to the query cell clusters, or defining a query cluster as an unassigned "novel" cell type not found in the reference.

Here, we extend the FR-Match algorithm to map each query *cell* to the known cell type clusters in the reference, i.e., cell-to-cluster matching. The input data are the same as before. If the query clusters are unavailable, it is sufficient to obtain broadly-grouped clusters by using the popular Louvain [17] or Leiden [18] clustering algorithms for scRNA-seq data. These clusters may not be at the ideal level of granularity to be directly matched to the granular cell types defined in the reference; rather, they provide candidate cluster memberships as the input data to FR-Match.

The extended cell-to-cluster FR-Match algorithm is implemented in the function `FRmatch_cell2cluster()`, and its plotting function implemented as `plot_FRmatch_cell2cluster()` in the `FRmatch` R package. The steps of the algorithm and the corresponding arguments in the functions are as follows:

1. Dimensionality reduction:
    - 1.1. Select informative marker genes using the companion marker gene selection algorithm - NS-Forest - or user-defined marker genes for the reference dataset;
    - 1.2. Extract the expression data for the reference marker genes in the query dataset, i.e., project the query data into the reference feature space for reduced dimensionality;
2. Pairwise iterative matching:
    - 2.1. For each pair of query cluster (j) and reference cluster (k):
        - 2.1.1. For subsample iteration index i iterating from 1 to the total number of iterations (`subsamp.iter=`):
            - 2.1.1.1. Subsample the same number of cells (`subsamp.size=`) from the query and reference clusters, denoted as S_i for the set of selected query cells;
            - 2.1.1.2. Perform Friedman-Rafsky test (FR test) [19], a nonparametric statistical test for multivariate two-group comparison, and obtain p-value from the test, denoted as p_i;
            - 2.1.1.3. Assign the p-value to the selected query cells, i.e., $p_{ck} = p_i$ for $c \in S_i$ and reference cluster k;
            - 2.1.1.4. Repeat 2.1.1.1 and 2.1.1.2, and obtain $p_{i'}$ for the updated iteration i';
            - 2.1.1.5. Update $p_{ck} = \max\{p_{ck}, p_{i'}\}$ for $c \in S_{i'}$ and reference cluster k, i.e., re-assign p_{ck} if $p_{i'}$ is greater than previously assigned p_{ck};
        - 2.1.2. End looping over iterations;
    - 2.2. End looping over query-and-reference-cluster-pairs;
    - 2.3. Obtain a p-value matrix $\{p_{ck}\}$ for every query cell c and reference cluster k;
    - 2.4. Apply multiple hypothesis testing correction to the p-values (`p.adj.method=`);

2.5. Determine the matched cell type for a query cell as the reference cell type that gives the maximum p-value for that query cell, or unassigned (i.e., no matched cell type) if the maximum p-value is below the p-value threshold (`sig.level=`).

Normalization

The plate-based SMART-seq and droplet-based 10X Genomics protocols are known to have very different read count distributions and detection sensitivities [20]. Thus, normalization is a key step for performing matching across these platforms. In our pipeline, we designed a rescaling and normalization procedure based on the expression value distributions and the signal-to-background-noise patterns observed in the cell type barcode plots.

First, we observed that the gene expression values of the SMART-seq and 10X data had very different dynamic range (Supplementary Figure 1). The marker genes displayed in the cell type barcode were selected by the NS-Forest marker gene selection algorithm that preferentially selects binary expression genes [12], i.e., those genes that are highly expressed in the given cell type and have no/weak expression in other cell types. For the purpose of cross-platform comparison, we designed a gene-wise min-max rescaling step to align the dynamic range of gene expression of both protocols in the range of [0,1]. Let $\boldsymbol{x}_g$ be a length-N vector of the expression value of marker gene g across all N cells in the dataset. The rescaled expression vector is

$$\tilde{\boldsymbol{x}}_g = \frac{\boldsymbol{x}_g}{\max(\boldsymbol{x}_g)}.$$

Second, due to the higher sensitivity of the SMART-seq protocol and low detection rate of the 10X protocol for weakly expressed genes, the cell type barcodes displayed weak signals for the genes that are not the marker genes of the given cell type in the SMART-seq data, whereas the cell type barcode of the 10X data displayed zero expression for those genes. For the purpose of cell type matching, the weak expression in the SMART-seq cell type barcodes can be considered a kind of background noise in its expression pattern (Figure 1A). In order to eliminate such background noise in the SMART-seq barcode, we designed the following normalization step. Let $\tilde{X}_b$ be the rescaled but unnormalized expression sub-matrix displayed in a cell type barcode b. $\tilde{X}_b$ is an $m \times n_b$ matrix, where m is the number of all marker genes, and n_b is the number of cells of cell type b. The normalized values are

$$X_b^{normalized} = \boldsymbol{w}_b \cdot \tilde{X}_b$$

where $\boldsymbol{w_b}$ is a weighting vector consisting of the row means (or medians) of $\tilde{X}_b$. Due to the binaryness of NS-Forest marker genes, $\boldsymbol{w}_b$ is usually a binary vector with values either close-to-0 or close-to-1. Due to the weighting, the dynamic range of the normalized values may shrink from [0,1]. A final rescaling step is to realign the maximum value of the dynamic range back to 1 sub-matrix-wise, which is:

$$X_b^{final} = \frac{1}{\max(X_b^{normalized})} \cdot X_b^{normalized}.$$

The final expression matrix for the input of the algorithm is the column-concatenation of X_b^{final} for all b's, where $N = \sum_b n_b$.

The above procedure is implemented in the normalization function `normalization()` in the R package. In the matching use cases presented, the weighting normalization procedure was only applied in the case of cross-platform matching between SMART-seq and 10X protocols. If both the query and reference data are generated using the same platform, the weighting step is not necessary, which can

be turned on or off by specifying `norm.by="mean"`, `norm.by="median"`, or `norm.by=NULL` options in the `normalization()` function.

Cosine distance metric in FR-Match

To make more robust matching, we made another modification in the FR-Match algorithm, which is to calculate the cosine distance that is invariant to scaling as an option for constructing the minimum spanning tree used in the FR test. Let $\boldsymbol{x} = (x_g)_{g=1}^{m}$ and $\boldsymbol{y} = (y_g)_{g=1}^{m}$ be two cells in the m-dimensional feature space of marker genes $g = 1, \cdots, m$. The cosine similarity between the two cells is defined as

$$\text{similarity} = \cos(\theta) = \frac{\sum_{g=1}^{m} x_g \cdot y_g}{\sqrt{\sum_{g=1}^{m} x_g^2} \cdot \sqrt{\sum_{g=1}^{m} y_g^2}}$$

where θ is the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Intuitively, if the angle θ is small, then $\cos(\theta)$ is large, which means the two cells $\boldsymbol{x}$ and $\boldsymbol{y}$ are more similar to each other as the angle between their representing vectors is small in the multi-dimensional space. If two cells are from different platforms, say $\boldsymbol{x}$ is SMART-seq data and $\boldsymbol{y}$ is 10X data, the scale difference between their expression range is normalized by the denominator in the above equation, which is the product of the lengths of the two vectors. Finally, the cosine distance is defined as

$$\text{distance} = 1 - \cos(\theta).$$

It is suggested to use the scaling-invariant cosine distance for more robust cell type matching across platforms. The option of using cosine distance can be turned on or off by specifying `use.cosine=TRUE` in the `FRmatch()` or `FRmatch_cell2cluster()` functions.

Data availability

All datasets are publicly available in Allen Brain Map Cell Types Database: RNA-Seq Data (https://portal.brain-map.org/) and NeMO Data Archieve (https://nemoarchive.org/). Specifically, each dataset can be downloaded from the following list.

- Human M1 10X: https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x
- Human M1 SMART-seq: https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq
- Mouse MOp single-nucleus RNA-seq: https://assets.nemoarchive.org/dat-ch1nqb7
- Mouse MOp single-cell RNA-seq: https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-10x
- Human MTG SMART-seq: https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq
- Mouse VISp single-cell RNA-seq and smFISH: https://portal.brain-map.org/atlases-and-data/rnaseq/data-files-2018

Raw count matrices were downloaded and preprocessed by log-transformation of the count per million (CPM) data. Log(CPM) data were the input data of the FR-Match algorithm.

Code availability

Open source software packages – NS-Forest and FR-Match – are available in GitHub repositories. Reproducible analysis notebooks are also available as tutorials in the software GitHub page. All details can be found in https://jcventerinstitute.github.io/celligrate/.

Acknowledgements

The work reported in this manuscript was funded by the JCVI Innovation Fund, the Allen Institute for Brain Science, and the U.S. National Institutes of Health (1RF1MH123220). The funding bodies had no role in the design or conclusions of this study.

Author contributions

YZ and RS conceived the project and prepared the manuscript. YZ and BA conducted the analyses. RA identified and provided the datasets. All authors agreed on the manuscript.

Competing interests

None

Figure Legends

Figure 1: Cell type matching based on features selected using NS-Forest and feature distribution matching using FR-Match. A. Schematic of the cross-platform cell type matching pipeline. The pipeline includes input reference data from the 10X Genomics platform, marker gene selection using NS-Forest algorithm based on the reference gene expression data, marker gene expression extraction for the query input SMART-seq platform data using the 10X reference marker genes, platform-specific rescaling and normalization steps, and marker gene expression distribution matching of the query and reference normalized data using the FR-Match algorithm. **B.** Cell-to-cluster matching results for matching cell types from SMART-seq (query) to 10X (reference) datasets for the human M1 brain region using FR-Match. Results are shown as the proportion of cells matched between pairs of query and reference subclass cell types. Most of the query cells are matched with the expected reference cell type subclass, aligning diagonally in the plot. The only exception is the agglomerated query IT subclass that was matched to several layer-specific reference IT subclasses or unassigned. **C.** Cell-to-cluster matching results for matching cell types from single-cell RNA-seq (scRNA-seq) (query) to single-nucleus RNA-seq (snRNA-seq) (reference) 10X datasets for mouse MOp brain region using FR-Match. Highlighted in the red box is an example of evidence for an under-partitioned query SMC-Peri subclass in which cells were matched to either the SMC subclass or the Peri subclass in the reference dataset. **D.** Cell type barcodes of the query SMC-Peri subclass and the corresponding reference SMC subclass and reference Peri subclass. The barcode plots clearly show two distinct expression patterns in the query cluster, each reflecting one of the two reference cluster expression pattern, supporting the under-partitioning hypothesis.

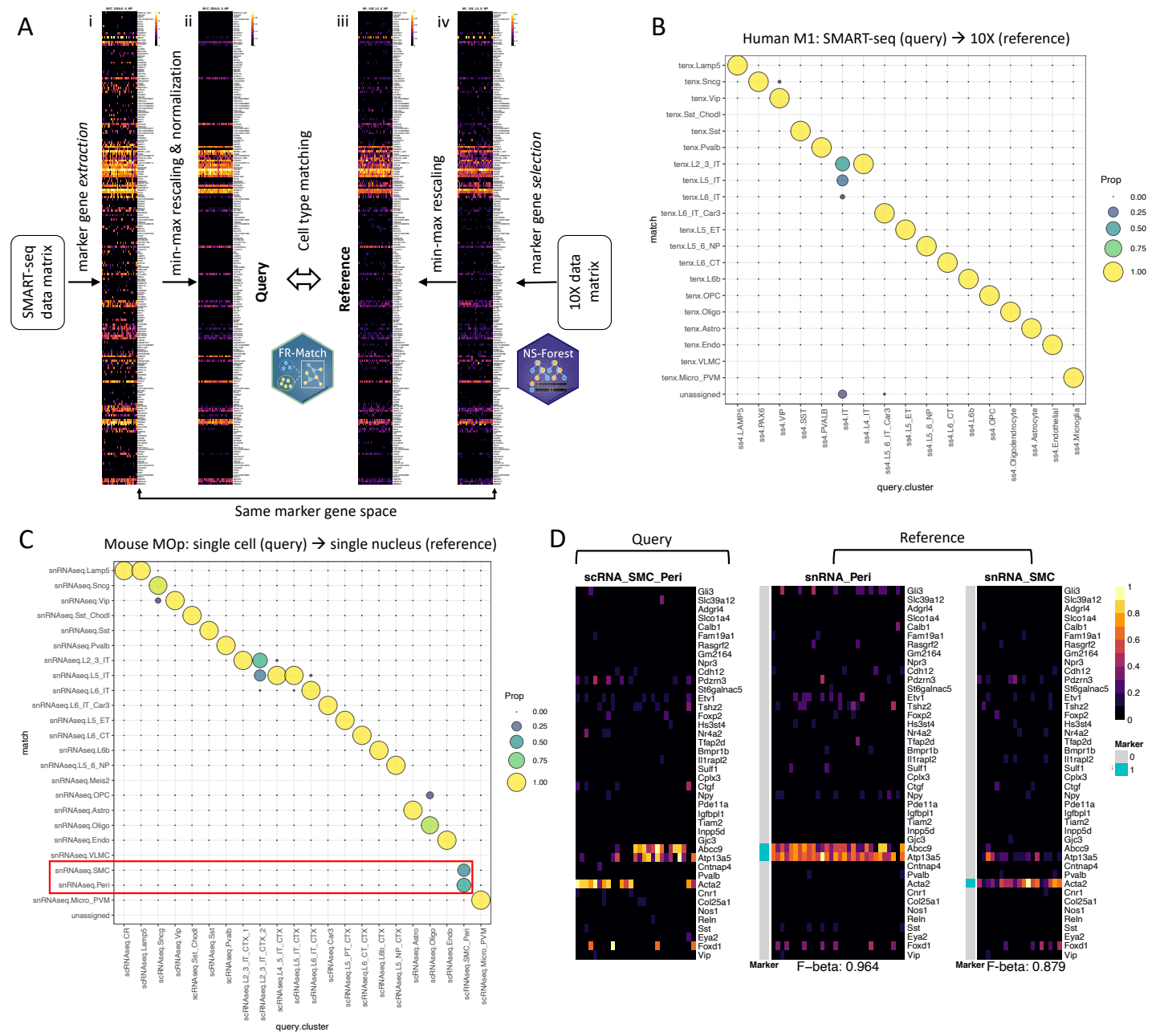
Figure 2: Additional FR-Match use cases and comparisons. A. Cell-to-cluster matching results for matching cell types from scRNA-seq (query) to snRNA-seq (reference) datasets of mouse MOp cell types at the most granular cell type resolution using FR-Match. The majority of cells in the query cell types were matched uniquely to reference cell types, showing clean diagonal matches with few off-diagonal matches. The highlighted box (in red) at the bottom is the "unassigned" row for the query cells that were not matched to any of the reference cell types based on the FR-Match results. The unassigned cells may correspond to novel query cell types not presented in the reference cell types. **B.** Matching results for matching cell types from scRNA-seq (query) to snRNA-seq (reference) datasets of mouse MOp cell type matching at the most granular cell type resolution using Azimuth. Though the majority of the cells were matched along the diagonal, there were many off-diagonal matches suggesting lower-quality matching. **C.** Cluster-to-cluster two-way matching results for matching cell types across the human M1 and MTG brain regions using FR-Match. FR-Match results suggests that most of the GABAergic and glial cell types are conserved across brain regions, whereas the glutamatergic cell types appear to be region-specific. **D.** Application of FR-Match to the spatial transcriptomics data. Cell type assignment of a mouse VISp smFISH dataset using scRNA-seq-defined reference cell types of the same brain region and the FR-Match cell-to-cluster algorithm. The assigned excitatory cell types clearly recapitulate the laminar distribution in the spatial coordinates (top); the assigned inhibitory cell types show the expected scattered spatial distribution. **E.** Spatial distributions of the excitatory cell types (top) and inhibitory cell types (bottom) summarized from the FR-Match cell type assignment results for the smFISH data shown in D.

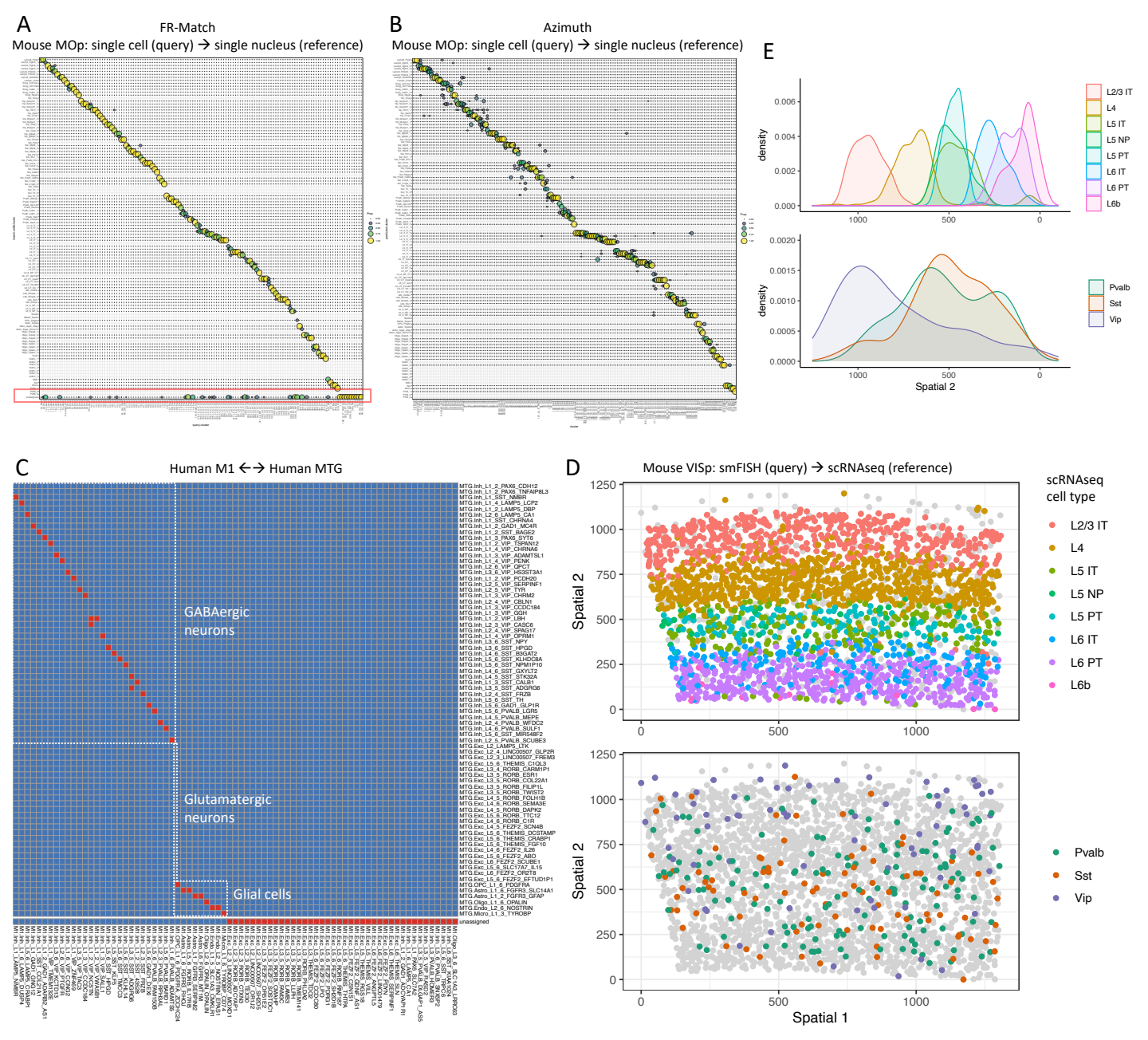
Supplementary Figure 1: SMART-seq and 10X data distributions. Histograms of the log(CPM) data from the SMART-seq protocol (left) and the 10X protocol (right) are shown. The SMART-seq data form a bimodal distribution, whereas the 10X data form a long-tail right-skewed distribution.

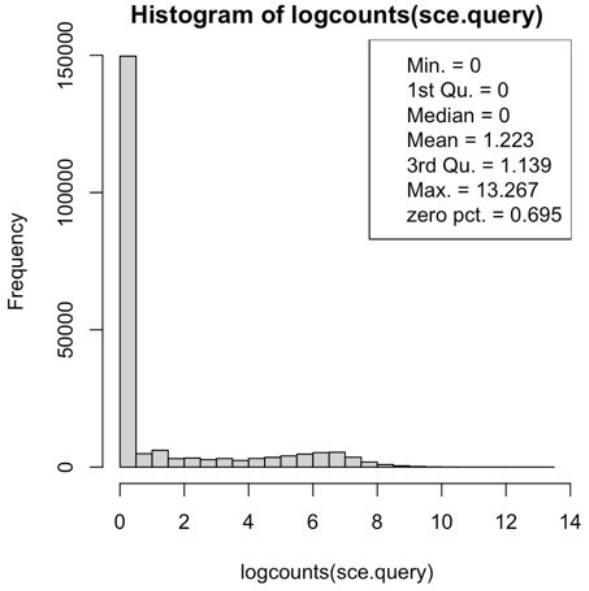
Supplementary Figure 2: Cell type matching results using Azimuth and Online iNMF. A. SMART-seq (query) to 10X (reference) matching results of human M1 subclass cell types using Azimuth. **B.** SMART-seq (query) to 10X (reference) matching results of human M1 subclass cell types using Online iNMF. **C.** ScRNA-seq (query) to snRNA-seq (reference) matching results of mouse MOp subclass cell types using Azimuth. Highlighted in the red box is the potentially under-partitioned query SMC-Peri subclass. **D.** ScRNA-seq (query) to snRNA-seq (reference) matching results of mouse MOp subclass cell types using Online iNMF. Highlighted in the red box is the potentially under-partitioned query SMC-Peri subclass.

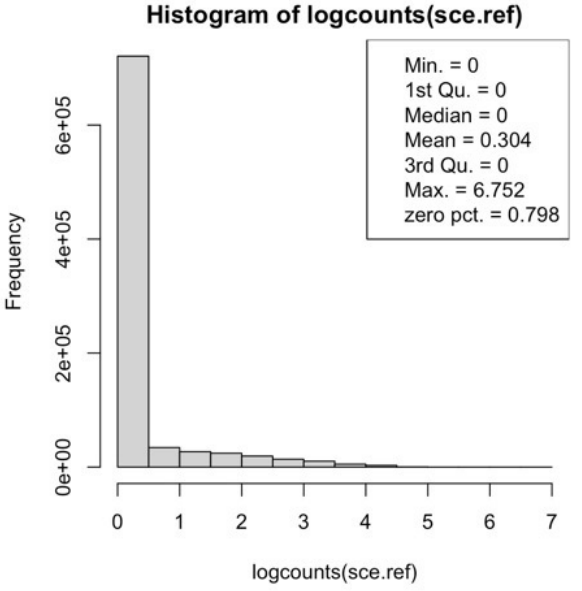
Supplementary Figure 3: Matching results for matching cell types from SMART-seq (query) to 10X (reference) datasets of human M1 cell types at the most granular cell type resolution using FR-Match (left) or Azimuth (right).

Supplementary Figure 4: Cell Type Barcode of reciprocal marker genes. Barcode plots for matched cell types (M1.Inh_L1_LAMP5_NMBR and MTG.Inh_L1_SST_NMBR) between M1 and MTG. Matched cell types were identified by two-way FR-Match. Blue vertical bars highlight the marker genes selected for the cell types displayed. The F-beta scores of classification accuracy using the marker gene combinations are listed at the bottom. Left pair are barcodes of the two cell types based on the marker genes derived using the M1 dataset. Right pair are barcodes of the two cell types based on the marker genes derived using the MTG dataset. Both pairs show very similar expression patterns of the barcodes within each pair on the reciprocal sets of marker genes, supporting the close similarity of the matched cell types between the different brain regions.

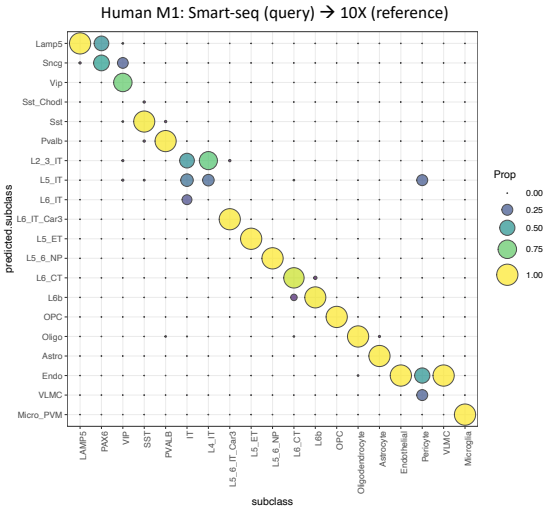

A
i
ii
iii
iv
SMART-seq data matrix
marker gene extraction
min-max rescaling & normalization
Query
Cell type matching
Reference
min-max rescaling
marker gene selection
10X data matrix
FR-Match
NS-Forest
Same marker gene space
B
Human M1: SMART-seq (query) → 10X (reference)
tenx.Lamp5
tenx.Sncg
tenx.Vip
tenx.Sst_Chodl
tenx.Sst
tenx.Pvalb
tenx.L2_3_IT
tenx.L5_IT
tenx.L6_IT
tenx.L6_IT_Car3
tenx.L5_ET
tenx.L5_6_NP
tenx.L6_CT
tenx.L6b
tenx.OPC
tenx.Oligo
tenx.Astro
tenx.Endo
tenx.VLMC
tenx.Micro_PVM
unassigned
match
ss4.LAMP5
ss4.PAX6
ss4.VIP
ss4.SST
ss4.PVALB
ss4.IT
ss4.L4_IT
ss4.L5_6_IT_Car3
ss4.L5_ET
ss4.L5_6_NP
ss4.L6_CT
ss4.L6b
ss4.OPC
ss4.Oligodendrocyte
ss4.Astrocyte
ss4.Endothelial
ss4.Microglia
query.cluster
Prop
0.00
0.25
0.50
0.75
1.00
C
Mouse MOp: single cell (query) → single nucleus (reference)
snRNAseq.Lamp5
snRNAseq.Sncg
snRNAseq.Vip
snRNAseq.Sst_Chodl
snRNAseq.Sst
snRNAseq.Pvalb
snRNAseq.L2_3_IT
snRNAseq.L5_IT
snRNAseq.L6_IT
snRNAseq.L6_IT_Car3
snRNAseq.L5_ET
snRNAseq.L6_CT
snRNAseq.L6b
snRNAseq.L5_6_NP
snRNAseq.Meis2
snRNAseq.OPC
snRNAseq.Astro
snRNAseq.Oligo
snRNAseq.Endo
snRNAseq.VLMC
snRNAseq.SMC
snRNAseq.Peri
snRNAseq.Micro_PVM
unassigned
match
scRNAseq.CR
scRNAseq.Lamp5
scRNAseq.Sncg
scRNAseq.Vip
scRNAseq.Sst_Chodl
scRNAseq.Sst
scRNAseq.Pvalb
scRNAseq.L2_3_IT_CTX_1
scRNAseq.L2_3_IT_CTX_2
scRNAseq.L4_5_IT_CTX
scRNAseq.L5_IT_CTX
scRNAseq.L6_IT_CTX
scRNAseq.Car3
scRNAseq.L5_PT_CTX
scRNAseq.L6_CT_CTX
scRNAseq.L6b_CTX
scRNAseq.L5_NP_CTX
scRNAseq.Astro
scRNAseq.Oligo
scRNAseq.Endo
scRNAseq.SMC_Peri
scRNAseq.Micro_PVM
query.cluster
Prop
0.00
0.25
0.50
0.75
1.00
D
Query
Reference
scRNA_SMC_Peri
snRNA_Peri
snRNA_SMC
Gli3
Slc39a12
Adgrl4
Slco1a4
Calb1
Fam19a1
Rasgrf2
Gm2164
Npr3
Cdh12
Pdzrn3
St6galnac5
Etv1
Tshz2
Foxp2
Hs3st4
Nr4a2
Tfap2d
Bmpr1b
Il1rapl2
Sulf1
Cplx3
Ctgf
Npy
Pde11a
Igfbpl1
Tiam2
Inpp5d
Gjc3
Abcc9
Atp13a5
Cntnap4
Pvalb
Acta2
Cnr1
Col25a1
Nos1
Reln
Sst
Eya2
Foxd1
Vip
Marker
F–beta: 0.964
F-beta: 0.879
1
0.8
0.6
0.4
0.2
0
Marker
0
1


A
FR-Match
Mouse MOp: single cell (query) → single nucleus (reference)
B
Azimuth
Mouse MOp: single cell (query) → single nucleus (reference)
E
L2/3 IT
L4
L5 IT
L5 NP
L5 PT
L6 IT
L6 PT
L6b
density
Pvalb
Sst
Vip
Spatial 2
C
Human M1 ←→ Human MTG
GABAergic neurons
Glutamatergic neurons
Glial cells
D
Mouse VISp: smFISH (query) → scRNAseq (reference)
scRNAseq cell type
Spatial 2
Spatial 1

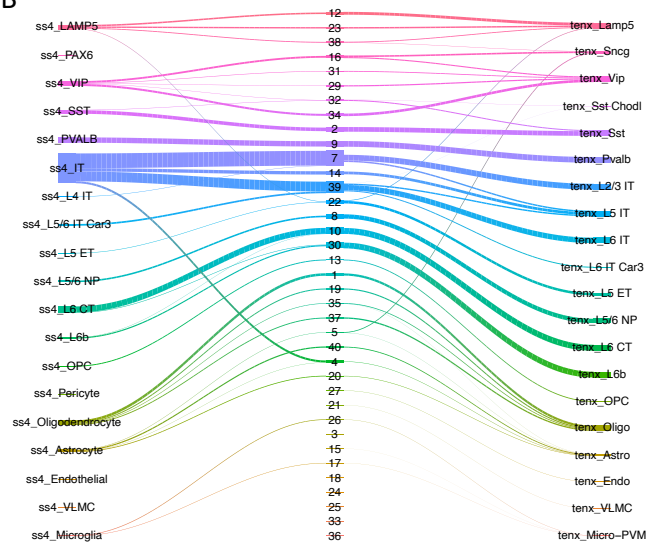

Histogram of logcounts(sce.query)
Min. = 0
1st Qu. = 0
Median = 0
Mean = 1.223
3rd Qu. = 1.139
Max. = 13.267
zero pct. = 0.695
Frequency
0
50000
100000
150000
0
2
4
6
8
10
12
14
logcounts(sce.query)


Histogram of logcounts(sce.ref)
Min. = 0
1st Qu. = 0
Median = 0
Mean = 0.304
3rd Qu. = 0
Max. = 6.752
zero pct. = 0.798
Frequency
0e+00
2e+05
4e+05
6e+05
0
1
2
3
4
5
6
7
logcounts(sce.ref)

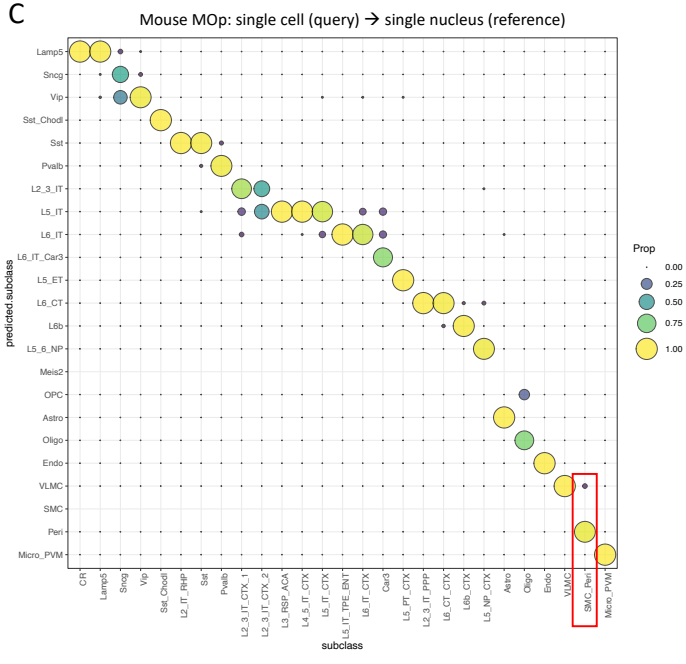
A


Human M1: Smart-seq (query) → 10X (reference)
predicted.subclass
subclass
Prop
0.00
0.25
0.50
0.75
1.00
Lamp5
Sncg
Vip
Sst_Chodl
Sst
Pvalb
L2_3_IT
L5_IT
L6_IT
L6_IT_Car3
L5_ET
L5_6_NP
L6_CT
L6b
OPC
Oligo
Astro
Endo
VLMC
Micro_PVM
LAMP5
PAX6
VIP
SST
PVALB
IT
L4_IT
L5_6_IT_Car3
L5_ET
L5_6_NP
L6_CT
L6b
OPC
Oligodendrocyte
Astrocyte
Endothelial
Pericyte
VLMC
Microglia

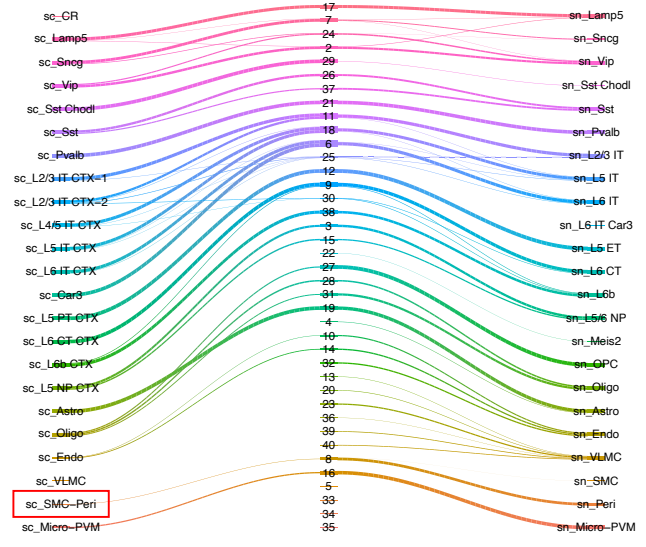
B

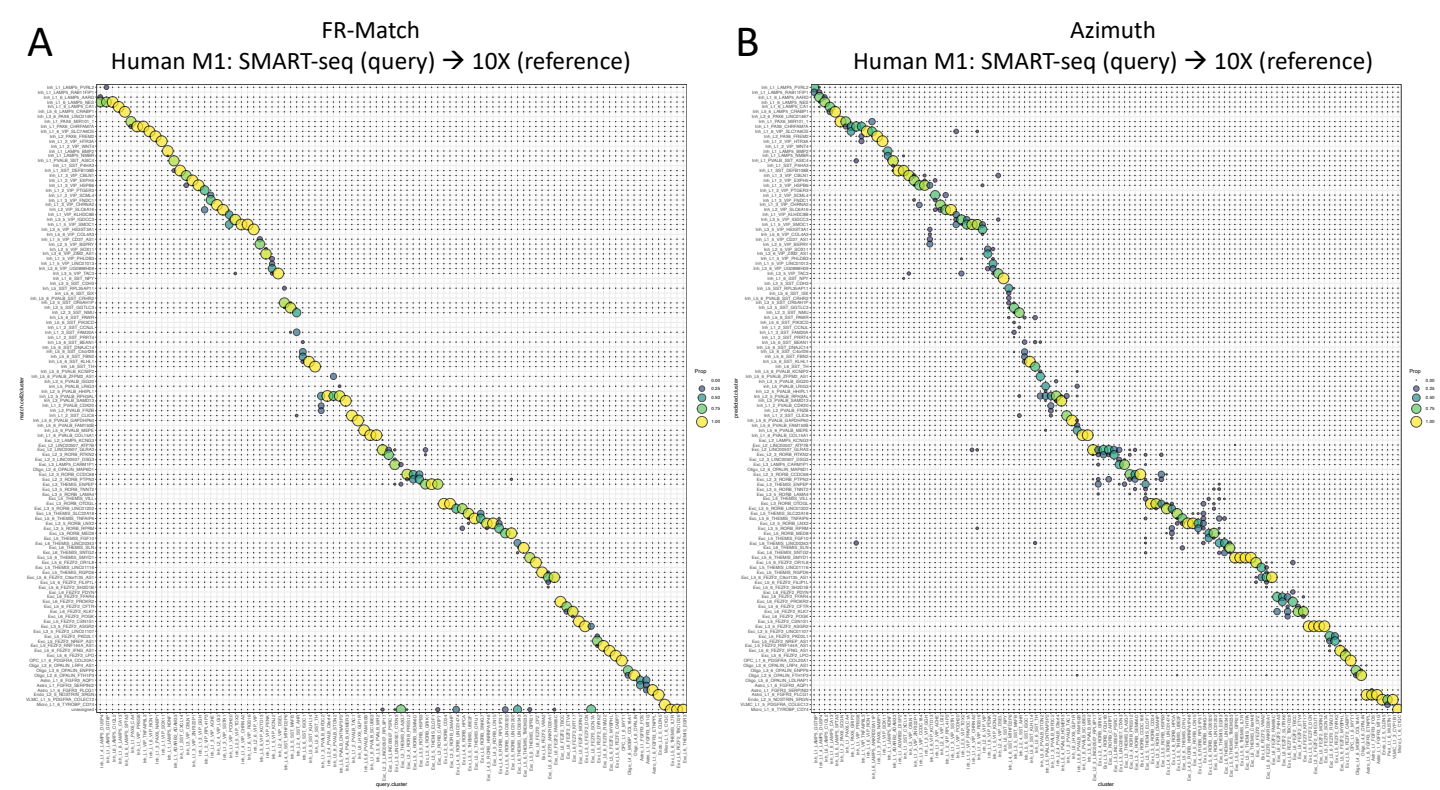

ss4_LAMP5
ss4_PAX6
ss4_VIP
ss4_SST
ss4_PVALB
ss4_IT
ss4_L4 IT
ss4_L5/6 IT Car3
ss4_L5 ET
ss4_L5/6 NP
ss4_L6 CT
ss4_L6b
ss4_OPC
ss4_Pericyte
ss4_Oligodendrocyte
ss4_Astrocyte
ss4_Endothelial
ss4_VLMC
ss4_Microglia
12
23
38
16
31
29
32
34
2
9
7
14
39
22
8
10
30
13
1
19
35
37
5
40
4
20
27
21
26
3
15
17
18
24
25
33
36
tenx_Lamp5
tenx_Sncg
tenx_Vip
tenx_Sst Chodl
tenx_Sst
tenx_Pvalb
tenx_L2/3 IT
tenx_L5 IT
tenx_L6 IT
tenx_L6 IT Car3
tenx_L5 ET
tenx_L5/6 NP
tenx_L6 CT
tenx_L6b
tenx_OPC
tenx_Oligo
tenx_Astro
tenx_Endo
tenx_VLMC
tenx_Micro–PVM

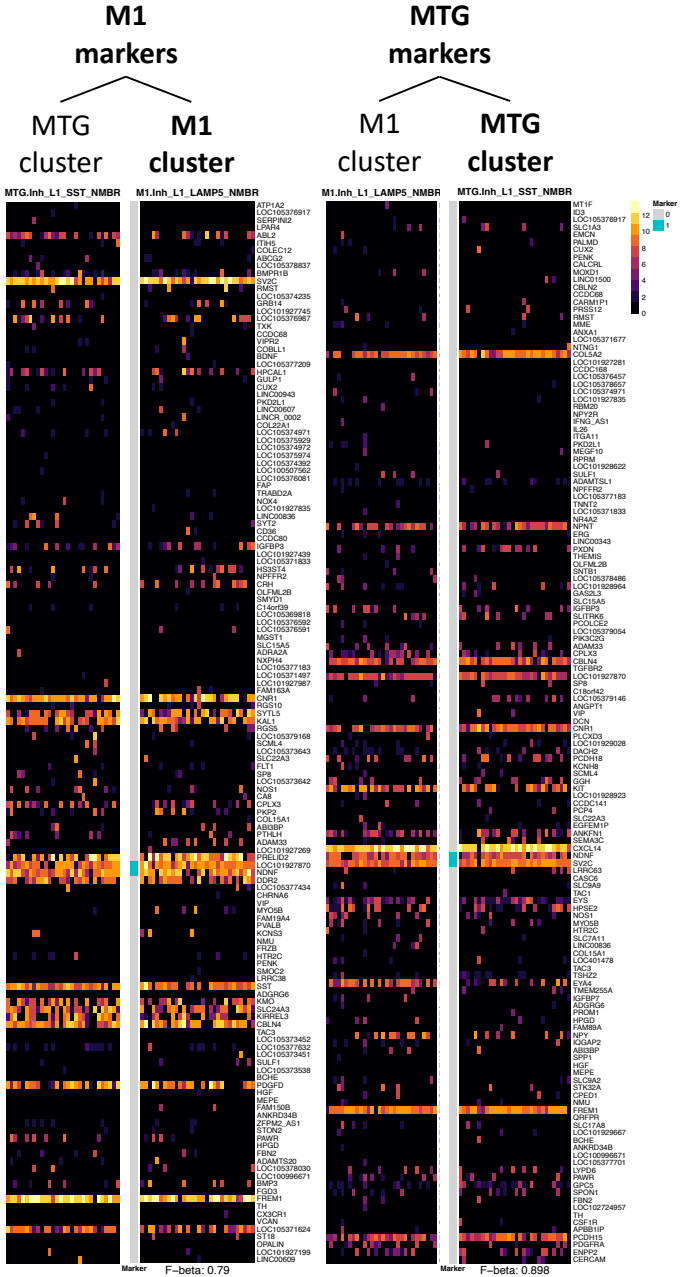
C


Mouse MOp: single cell (query) → single nucleus (reference)
predicted.subclass
subclass
Prop
0.00
0.25
0.50
0.75
1.00
Lamp5
Sncg
Vip
Sst_Chodl
Sst
Pvalb
L2_3_IT
L5_IT
L6_IT
L6_IT_Car3
L5_ET
L6_CT
L6b
L5_6_NP
Meis2
OPC
Astro
Oligo
Endo
VLMC
SMC
Peri
Micro_PVM
CR
Lamp5
Sncg
Vip
Sst_Chodl
L2_IT_PHP
Sst
Pvalb
L2_3_IT_CTX_1
L2_3_IT_CTX_2
L3_RSP_ACA
L4_5_IT_CTX
L5_IT_CTX
L5_IT_TPE_ENT
L6_IT_CTX
Car3
L5_PT_CTX
L2_3_IT_PPP
L6_CT_CTX
L6b_CTX
L5_NP_CTX
Astro
Oligo
Endo
VLMC
SMC_Peri
Micro_PVM

D


sc_CR
sc_Lamp5
sc_Sncg
sc_Vip
sc_Sst Chodl
sc_Sst
sc_Pvalb
sc_L2/3 IT CTX-1
sc_L2/3 IT CTX-2
sc_L4/5 IT CTX
sc_L5 IT CTX
sc_L6 IT CTX
sc_Car3
sc_L5 PT CTX
sc_L6 CT CTX
sc_L6b CTX
sc_L5 NP CTX
sc_Astro
sc_Oligo
sc_Endo
sc_VLMC
sc_SMC–Peri
sc_Micro–PVM
17
7
24
2
29
26
37
21
11
18
6
25
12
9
30
38
3
15
22
27
28
31
19
4
10
14
32
13
20
23
36
39
40
8
16
5
33
34
35
sn_Lamp5
sn_Sncg
sn_Vip
sn_Sst Chodl
sn_Sst
sn_Pvalb
sn_L2/3 IT
sn_L5 IT
sn_L6 IT
sn_L6 IT Car3
sn_L5 ET
sn_L6 CT
sn_L6b
sn_L5/6 NP
sn_Meis2
sn_OPC
sn_Oligo
sn_Astro
sn_Endo
sn_VLMC
sn_SMC
sn_Peri
sn_Micro–PVM


A
FR-Match
Human M1: SMART-seq (query) → 10X (reference)
query.cluster
Prop
B
Azimuth
Human M1: SMART-seq (query) → 10X (reference)
predicted.cluster
cluster
Prop


M1 markers
MTG cluster
M1 cluster
MTG.lnh_L1_SST_NMBR
M1.lnh_L1_LAMP5_NMBR
ATP1A2
LOC105376917
SERPINI2
LPAR4
ABL2
ITIH5
COLEC12
ABCG2
LOC105378837
BMPR1B
SVIL
RMST
LOC105374235
GRB14
LOC101927745
LOC105376987
TXK
CCDC68
VIPR2
COBLL1
BDNF
LOC105377209
HPCAL1
GULP1
CUX2
LINC00943
PKDCL1
LINC00607
LINC01_0002
COL22A1
LOC105374971
LOC105375929
LOC105374972
LOC105375974
LOC105374392
LOC100507562
LOC105376081
FAP
TRABD2A
NCX4
LOC101927835
LINC00836
SYT2
CD36
CCDC80
IGFBP3
LOC101927439
LOC105371833
HS3ST4
NPFFR2
CRH
OLFML2B
SMYD1
C14orf39
LOC105369818
LOC105376592
LOC105376591
MGST1
SLC15A5
ADRA2A
NXPH4
LOC105377183
LOC105371497
LOC101927987
FAM163A
CNR1
RGS10
SYTL5
KAL1
RGS5
LOC105379168
SCML4
LOC105373643
SLC22A3
FLT1
SP8
LOC105373642
NOS1
CA8
CPLX3
PKP2
COL15A1
ARG99P
PTHLH
ADAM33
LOC101927269
PRELID2
LOC101927870
NDNF
DDR2
LOC105377434
CHRNA6
VIP
MYO5B
FAM19A4
PVALB
KCNS3
NMU
FRZB
HTR2C
PENK
SMOC2
LRRC38
SST
ADGRG6
KMO
SLC24A3
KIRREL3
CBLN4
TAC3
LOC105373452
LOC105377632
LOC105373451
SULF1
LOC105373538
BCHE
PDGFD
HGF
MEPE
FAM150B
ANKRD34B
ZFPM2_AS1
STON2
PAWR
HPGD
FBN2
ADAMTS20
LOC105378030
LOC100996671
BMP3
FGD3
FREM1
TH
CX3CR1
VCAN
LOC105371624
ST18
OPALIN
LOC101927199
LINC00609
Marker
F−beta: 0.79
MTG markers
M1 cluster
MTG cluster
M1.lnh_L1_LAMP5_NMBR
MTG.lnh_L1_SST_NMBR
MT1F
ID3
LOC105376917
SLC1A3
EMCN
PALMD
CUX2
PENK
CALCRL
MOXD1
LINC01500
CBLN2
CCDC68
CARM1P1
PRSS12
RMST
MME
ANXA1
LOC105371677
NTN3
COL5A2
LOC101927281
CCDC168
LOC105376457
LOC105378657
LOC105374971
LOC101927835
RBM20
NPY2R
IFNG_AS1
IL26
ITGA11
PKDCL1
MEGF10
RPRM
LOC101928622
SULF1
ADAMTSL1
NPFFR2
LOC105377183
TNNT2
LOC105371833
NR4A2
NPNT
ERG
LINC00343
PXDN
THEMIS
OLFML2B
SNTB1
LOC105378486
LOC101928964
GAS2L3
SLC15A5
IGFBP3
SLITRK6
PCDLCE2
LOC105379054
PIK3C2G
ADAM33
CPLX3
CBLN4
TGFBR2
LOC101927870
SP8
C1qbr4q2
LOC105379146
ANGPT1
VIP
OCN
CNR1
PLCXD3
LOC101929028
DACH2
PCDH18
KCNH8
SCML4
SGIP1
KIT
LOC101928923
CCDC141
PCP4
SLC22A3
EGFEMIP
ANKFN1
SEMA3C
CXCL14
NDNF
SV2C
ERRC63
CASC9
SLC9A9
TAC1
EYS
HPSE2
NOS1
MYO5B
HTR2C
SLC7A11
LINC00836
COL15A1
LOC401478
TAC3
TSHZ2
EYA4
TMEM255A
IGFBP7
ADGRG6
PROM1
HPGD
FAM89A
NPY
IQGAP2
ABSBP
SPP1
HGF
MEPE
SLC9A2
STK32A
CPED1
NMU
FREM1
GRFP1
SLC17A8
LOC101929667
BCHE
ANKRD34B
LOC100996671
LOC105377701
LYPD6
PAWR
GPC5
SPON1
FBN2
LOC102724957
TH
CSF1R
APBB1P
PCDH15
PDGFRA
ENPP2
CERCAM
Marker
F−beta: 0.898
Marker
0
1
12
10
8
6
4
2
0