Identifying common and novel cell types in single-cell RNA-sequencing data using FR-Match

Authors: Yun Zhang[1], Brian Aevermann[1], Rohan Gala[2], Richard 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 novel cell types and non-optimally clustered cell types in the query data. This approach shows excellent performance for cross-platform, cross-sample type, cross-tissue region, and cross-data modality cell type matching.

Main

Single cell transcriptomic profiling has emerged as a powerful tool to survey and discover cell phenotype heterogeneity in complex biological systems. Large collaborative consortia, such as the Human Cell Atlas [1] and NIH BRIAN Initiative [2, 3], have widely adopted the unbiased single-cell/nucleus RNA-sequencing (scRNAseq) technologies to generate reference cell type atlases at single cell resolution across many organs and species. Transcriptomicly-defined cell types have uncovered cellular diversity at an unprecedented level of granularity; a series of recent publications have reported 128 transcriptomic 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 serving as a reference cell type database (https://portal.brain-map.org/atlases-and-data/rnaseq).

A key role for these reference datasets is to support the matching of new query data to the reference cell types. 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 effect, cell type matching is a more pragmatic use case that requires not only integrating the query cells onto 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 scRNAseq data including NS-Forest [11, 12] – a random forest machine learning algorithm for the identification of the minimum set of marker genes for given cell types and FR-Match [13] – a minimum spanning tree-based statistical testing approach for cell type matching of query and reference datasets. We introduced the concept of Cell Type Barcode [11, 13] using NS-Forest marker genes to visualize and characterize the distinction between different cell types [11]. The NS-Forest marker genes also serve as a dimensionality reduction approach for FR-Match that essentially matches the query and reference cell types based on the Cell Type Barcode 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 Barcode provides evidence and explainability for our matching results, as well as diagnostics of the cell type cluster quality.

We designed a normalization procedure based on the marker gene expression patterns observed in the Cell Type Barcode plots, to robustly remove technical artifacts observed in different scRNAseq platforms (Methods). 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 varying non-specific marker gene background expressions. The pre-normalization step is a min-max rescaling applied to each gene for both Smart-seq and 10X data, to globally align the data ranging from 0 to 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) as a single-value index of the expression level. The index vector is used to weight the per-Barcode expression pattern by multiplying to the per-Barcode expression matrix. Finally, the Smart-seq Barcode is again rescaled to [0,1] for matching. The above procedure effectively aligned the cross-platform Barcode patterns (Figure 1A(ii)(iii)), showing similar signal and noise levels.

The cell-to-cluster extension of FR-Match is 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 (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 (Methods).

Using the above-described pipeline and extensions, the cross-platform matching approach was validated using Allen human M1 snRNAseq data generated using the *10X Chromium v3* protocol [4] as the reference and an M1 snRNAseq 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 match of the subclass types for all query cells (Figure 1B), with the exception of the agglomerated IT query type. Due to the grouping of the layer-non-specific IT cells, the majority of these cells were matched to one of

the layer-specific IT reference types, with a few left unassigned. 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 *snRNAseq* from the Allen mouse MOp cell types [5] as the reference and *scRNAseq* from the MOp subset from a transcriptomic cell type taxonomy of the entire adult mouse isocortex and hippocampal formation [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 the subclass types, most of the query types were one-to-one matched to a reference type (Figure 1C). The highlighted box shows that the query SMC-Peri type was matched to the separate SMC and Peri types in the reference, with almost half-half split. As previously noted [13], a one-to-many match may suggest that the query cluster is under-partitioned. 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 reference Barcodes. 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 sub-optimal partitioning based on marker genes in the matching results.

We also benchmarked the FR-Match pipeline with the Azimuth and Online iNMF approaches (Supplementary Figure 2) for sub-optimal parititoning identification. All cells were matched, but these integration methods were not able to split the under-partitioned clusters. Azimuth matched all query SMC-Peri cells to the reference Peri subclass with 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 analyses, we also matched to the most granular cell types and benchmarked with Azimuth. For the mouse MOp cross-sample type matching results, FR-Match formed a clean diagonal alignment of cell types and assigned unmatched cells as “unassigned” in the bottom row (Figure 2A). Azimuth also matched the majority of the cells along the diagonal, but with many suboptimal matches scattered off-diagonally, which potentially should be unassigned (Figure 2B). Similar results for the human M1 cross-platform cell type matching 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 human MTG [6] brain regions and the bi-directional (M1 as query to MTG as reference, and vice versa) matching results 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 preserved across brain regions, whereas the excitatory neurons are region-

specific. We also examined the Cell Type Barcode plots for a pair of matched cell types (Supplementary Figure 4), showing highly similar expression patterns of the matched types using reciprocal marker genes, even though the best marker genes selected for each brain regions may be different.

Finally, we report the application of FR-Match for cross-data-modality cell type assignment using spatial transcriptomics data generated by single molecular fluorescence in situ hybridization (smFISH) [15] and Smart-seq scRNAseq as the reference, both from mouse primary visual cortex (V1) [16]. *De novo* clustering of the smFISH data was used to obtain broadly-defined clusters and the FR-Match cell-to-cluster pipeline to assign a reference cell type to each spatial cell in the initial clusters (Methods). The FR-Match results successfully recapitulated the clear laminar distributions of the excitatory neurons, corresponding to their assigned cell types (Figure 2D). In contrast, the inhibitory neurons are 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 fully reflected their location in the tissue, agreeing with the expected layering 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 vs. 10X), sample types (single-cell vs. single-nucleus), brain regions (M1 and MTG), and data modalities (spatial transcriptomics and scRNAseq). Compared with existing 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 underpinning 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., *Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse.* bioRxiv, 2020.
5. Yao, Z., et al., *An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types.* Biorxiv, 2020.
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., et al., *NS-Forest: A machine learning method for the objective identification of minimum marker gene combinations for cell type determination from single cell RNA sequencing.* bioRxiv, 2020.
12. 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.
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 graphic model and minimum spanning trees to learn the data distributional equivalence between two cell type clusters derived from single cell or single nucleus RNA sequencing (scRNAseq) data [13]. The required input data for FR-Match are cell-by-gene expression matrix 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 labels to the reference cluster labels, a.k.a. assigning the known reference cell types to the query cell clusters, or defining a query cluster as an unassigned "novel" cell type in the reference.

Here, we extend the FR-Match algorithm to map each query cell to the known cell types 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 scRNAseq 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 as follows, which is implemented in the function `FRmatch_cell2cluster()` and its plotting function `plot_FRmatch_cell2cluster()` in the `FRmatch` R package. Relevant arguments of the functions are also listed below.

1. Dimensionality reduction:
   1.1. Select informative marker genes using our companion marker gene selection algorithm - NS-Forest - or user-defined marker genes for the reference dataset;
   1.2. Extract the reference marker genes in the query dataset, i.e., project the query data to the reference feature space with reduced dimensionality;
2. Pairwise iterative matching:
   2.1. For each pair of query cluster (j) and reference cluster (k):
      2.1.1. For i in 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 protocol and the droplet-based 10X protocol are known to have very different read counts and sensitivity [20]. Normalization is a key step for performing the cross-platform matching. In our pipeline, we designed a rescaling and normalization procedure based on the expression values and the signal-and-background-noise-pattern observed from the Cell Type Barcode plots.

First, we observed that the gene expression values of the Smart-seq and 10X data have 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 [11], i.e., those genes that are highly expressed in the given cell type and have no/weak expression in other cell types. For the comparison purpose, 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 high sensitivity of Smart-seq protocol and low detection rate of 10X protocol on the weakly expressed genes, the Cell Type Barcode displays weak signal for the genes that are not the marker genes of the given cell type for the Smart-seq data, whereas the Cell Type Barcode of the 10X data displays zero expression for those genes. For the cell type matching purpose, the weak expression in the Cell Type Barcode of Smart-seq data becomes 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 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. The whole 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"` or `NULL` option 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 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 different 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()` function.

References (additional references)

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.

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

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

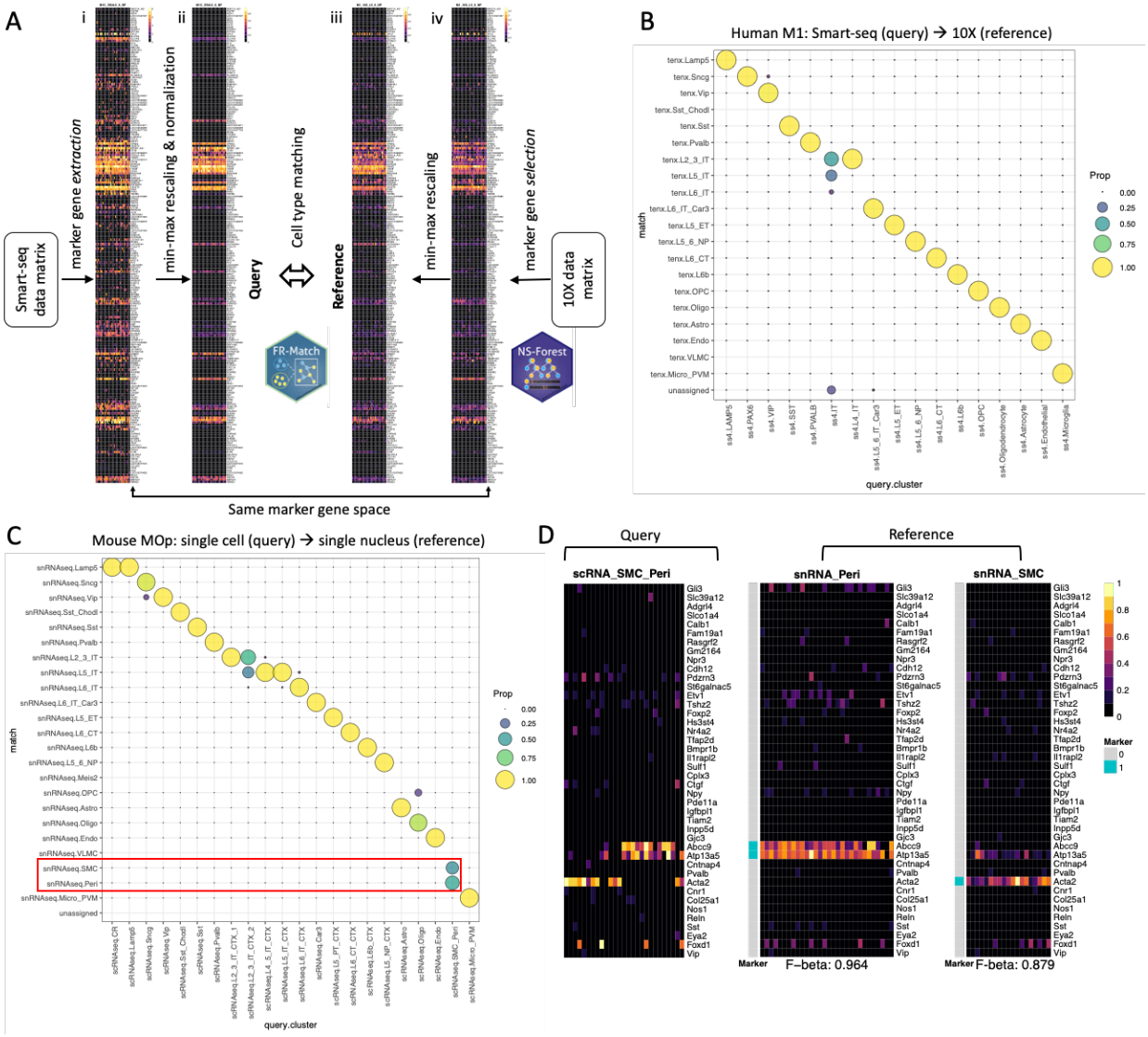


Figure 1: A. Schematic of cross-platform cell type matching pipeline based on feature selected using NS-Forest and cell type matching using FR-Match. The pipeline includes input reference data from the 10X platform, marker gene selection using NS-Forest algorithm based on the reference data, marker gene extraction for the query input Smart-seq platform data using the same set of reference marker genes, platform-specific rescaling and normalization steps, and cell type matching of the query and reference normalized data using FR-Match algorithm(s). **B.** FR-Match cell-to-cluster cell type matching results of the query Smart-seq data and reference 10X data for human M1 brain region. 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 subclass cell types, 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.** FR-Match cell-to-cluster cell type matching results of the query single-cell RNA-seq (scRNAseq) 10X data and reference single-nucleus RNA-seq (snRNAseq) 10X data for mouse MOp brain region. Highlighted boxed (in red) is the under-partitioned query SMC-Peri subclass that was split and matched to the corresponding SMC subclass and Peri subclass separately in the reference data. **D.** Cell Type Barcodes of the under-partitioned 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 under-partitioned query cluster, each reflecting a reference cluster expression pattern.

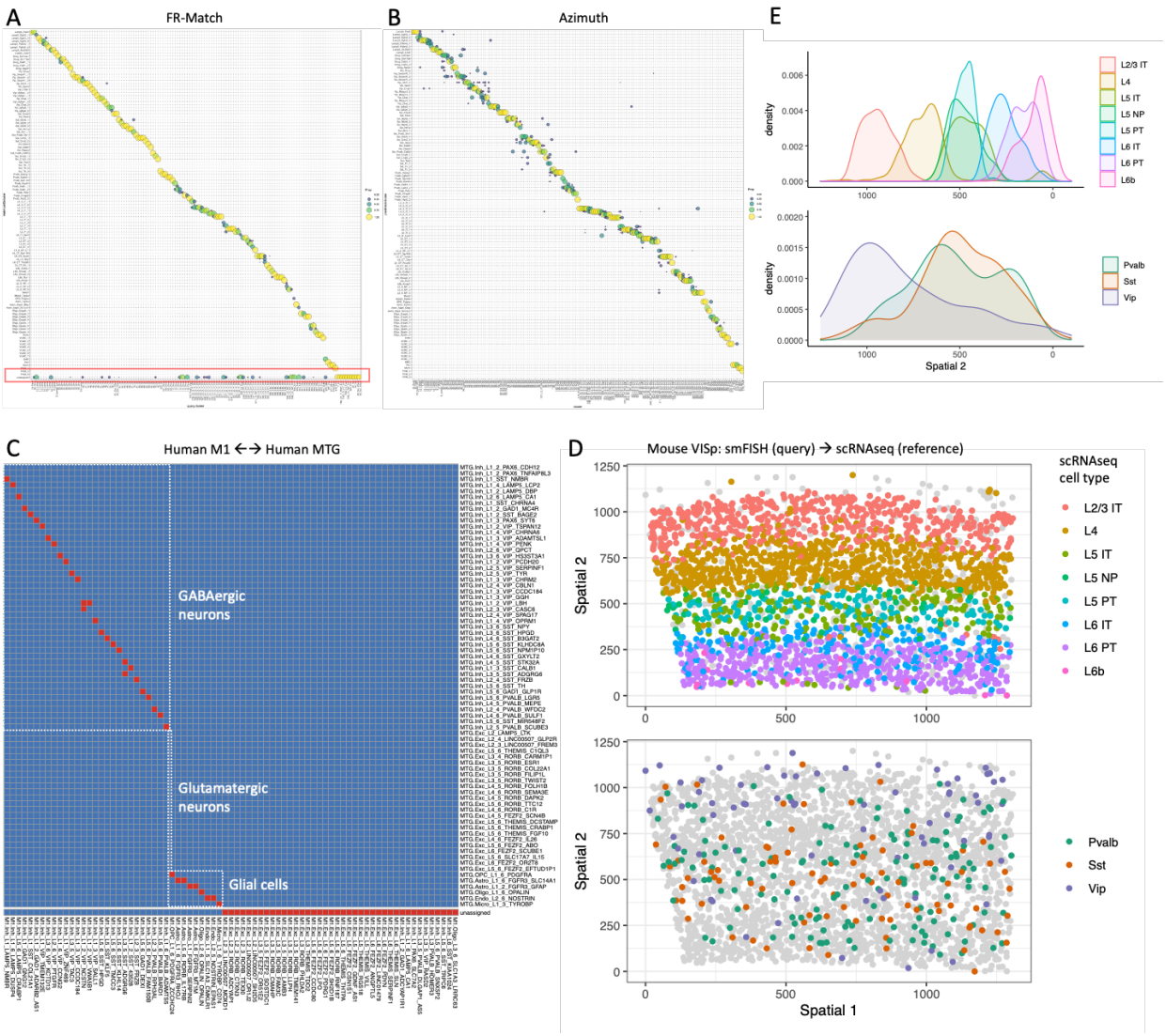
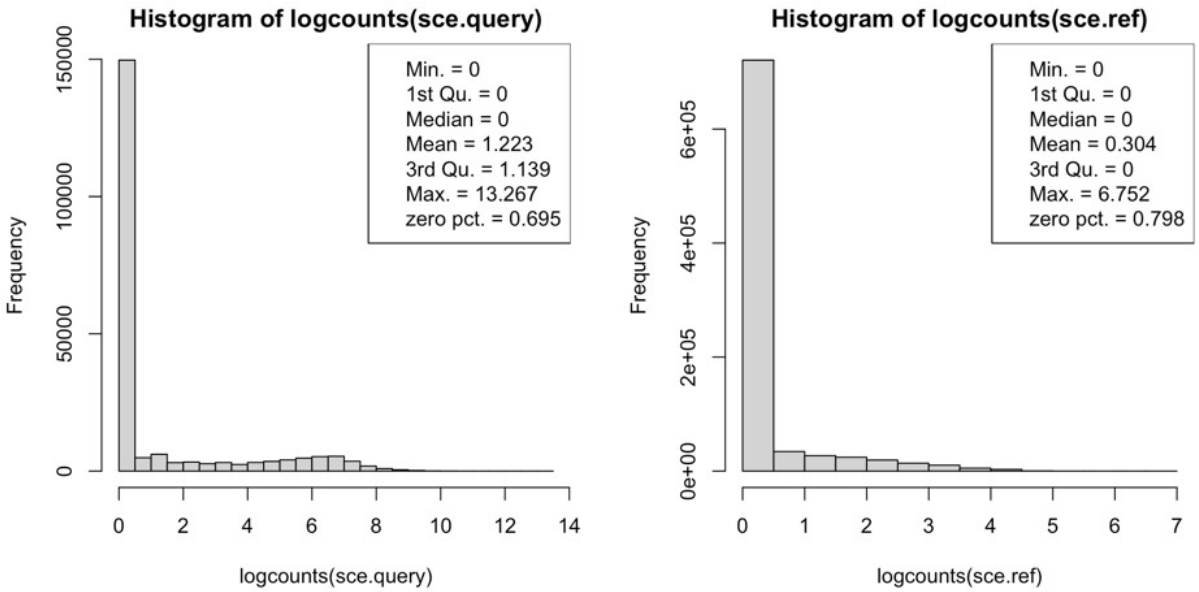
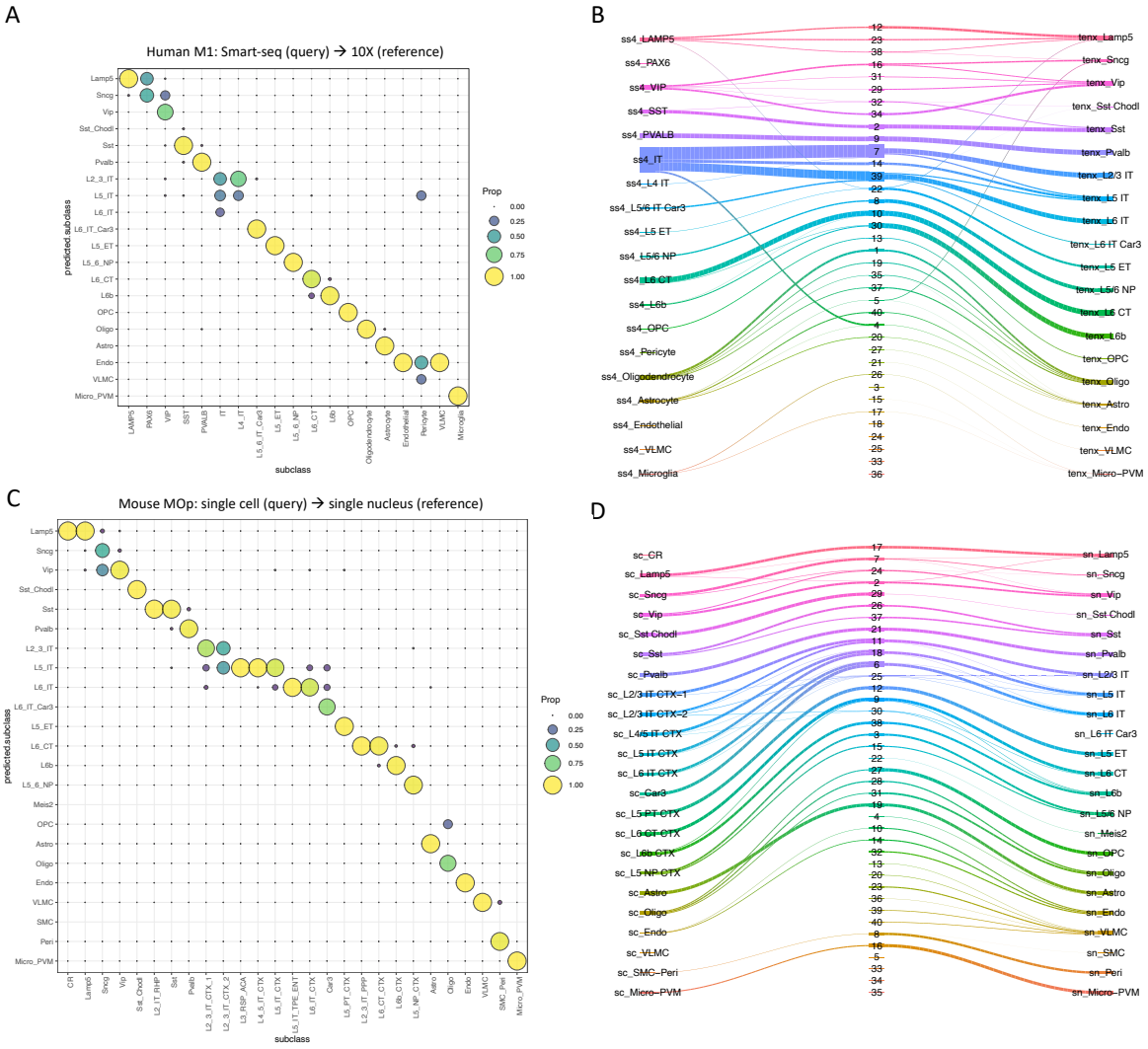


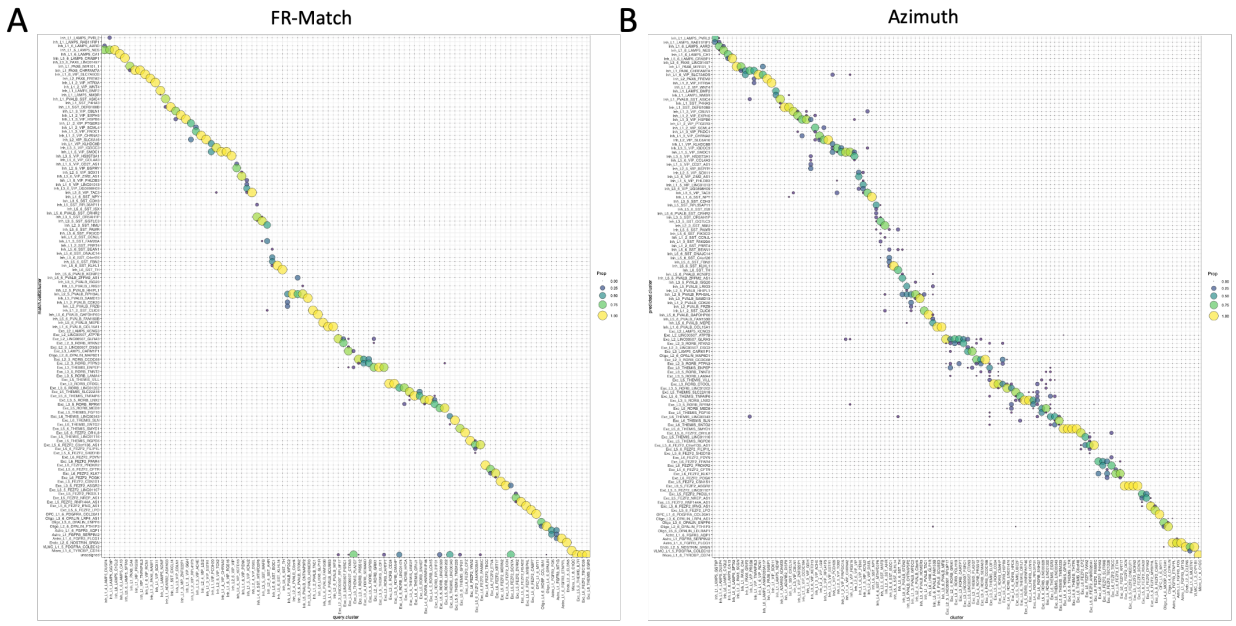
Figure 2: A. FR-Match cell-to-cluster results of mouse MOp cell type matching at the most granular cell type resolution. Majority of the cells in the query cell types were matched uniquely to the reference cell types, showing clean diagonal matches with few off-diagonal match. The highlighted box (in 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 restuls. The unassigned cells may suggest sub-optimally partitioned query clusters or novel cell types not presented in the reference cell types. **B.** Azimuth results of mouse MOp cell type matching at the most granular cell type resolution. Though majority of the cells were matched along the diagonal, there were many off-diagonal matches suggesting low-quality matching. **C.** FR-Match cluster-to-cluster two-way matching results for the cross-brain-region matching of human M1 and MTG. FR-Match results suggests that most of the GABAergic and glial cell types are preserved across brain regions, but Glutamatergic cell types are region-specific. **D.** FR-Match application to the spatial transcriptomics data. Cell type assignment of the mouse VISp smFISH data using the scRNAseq-defined reference cell types of the same brain region and FR-Match cell-to-cluster algorithm. The assigned excitatory cell types clearly recapitulate the laminar distribution in the spatial coordinates (top); and the assigned inhibitory cell types show the expected scattering 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.



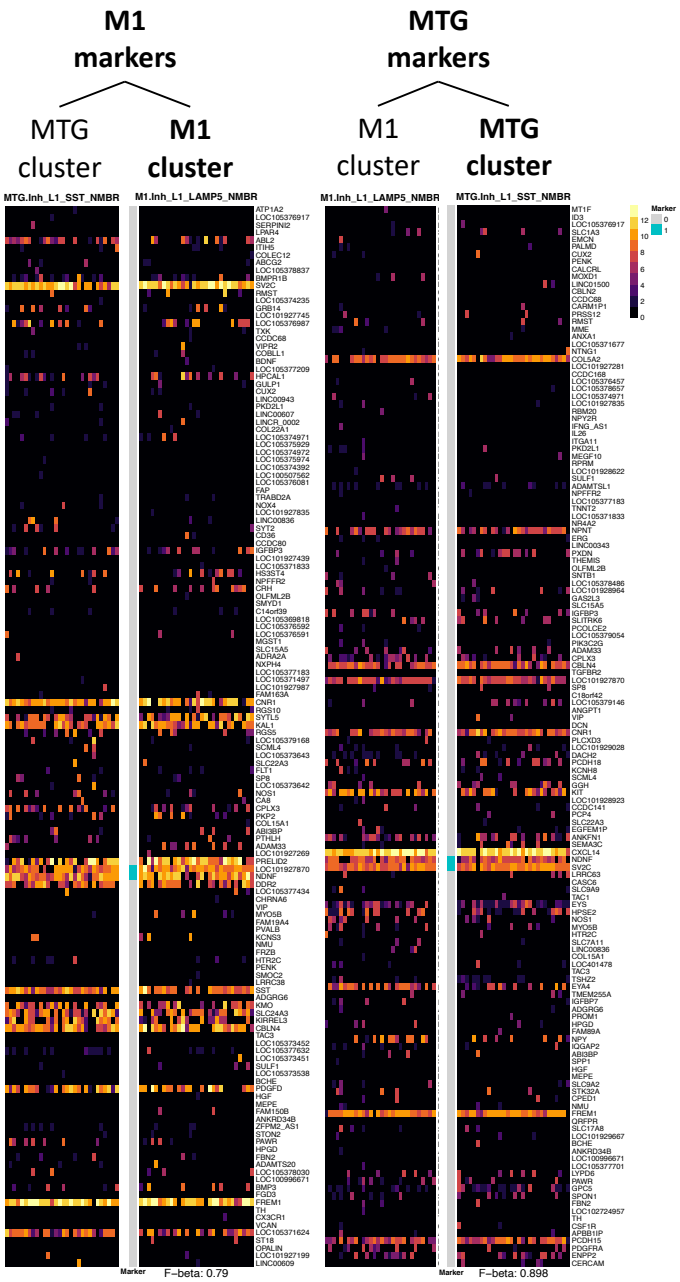
Supplementary Figure 1: Data distribution histogram of the log(CPM) data from the Smart-seq protocol (left) and the 10X protocol (right). The Smart-seq data form a bimodal distribution, whereas the 10X data form a long-tail right-skewed distribution.



Supplementary Figure 2: Cell type matching results using Azimuth and Online iNMF. A. Azimuth results of human M1 subclass cell type matching. **B.** Online iNMF results of human M1 subclass cell type matching. **C.** Azimuth results of mouse MOp subclass cell type matching. **D.** Online iNMF results of mouse MOp subclass cell type matching.



Supplementary Figure 3: FR-Match (left) and Azimuth (right) results of human M1 cell type matching at the most granular cell type resolution.



Supplementary Figure 4: Reciprocal marker gene Cell Type Barcodes. Pairs of Barcode plots for the 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. Left pair are Barcodes of the two cell types based on the M1 marker genes. Right pair are Barcodes of the two cell types based on the MTG marker genes. Both pairs show very similar expression patterns of the Barcodes within each pair on the reciprocal sets of marker genes, convincing the similarity of the matched cell types between different brain regions.