Optimizing single-cell RNA sequencing mapping algorithms for reference dataset construction

[1,2] Clare Morris, [1] Kelly Jin, [1] Changkyu Lee, [1] Zizhen Yao,
[1] Allen Institute for Brain Science, Seattle WA, [2] Harvard University, Cambridge MA

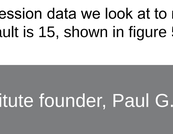
1. Introduction

- Single-cell RNA sequencing (scRNA-seq) provides whole transcriptome profiling at the individual cell level, revealing complex and rare cell type populations. These rare populations are best captured by large reference scRNA-seq datasets, in which millions of cells have been sequenced.
- Building a comprehensive scRNA-seq reference dataset requires mapping labels from a training to a test dataset. These training and test datasets are often of different modalities (i.e., droplet-based 10X Genomics Chromium (10X) vs. plate-based SmartSeq (SS) scRNA-seq platforms), raising concerns about differences in mapping quality between train and test daatasets.
- Mapping algorithms attempt to minimize the effects of different modalities to create more accurate reference datasets. We have performed a through comparison of Allen Institute's existing mapping algorithms, Flat and HKNN, in order to determine the optimal algorithm and parameters for future work constructing our scRNA-seq reference datasets.
- We demonstrate the use of one of these mapping algorithms by mapping an external scRNA-seq oligodendrocyte dataset to an internal dataset characterizing aged and adult non-neuronal cells in the mouse brain.

2. Methods

1. Assign a training and test dataset. In order to determine mapping accuracy, both the training and test dataset must have existing cluster annotations.
2. Identify cell types by transferring labels from the training dataset to the query dataset

3. Determine error, precision, and recall rates at the neighborhood, subclass, and cluster levels to evaluate mapping algorithm accuracy.



Error rate: $\frac{\sum TP+TN}{\sum TP+FP+FN+TN}$

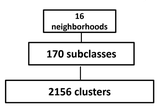
Recall: $\frac{\sum TP}{\sum TP+FN}$

Precision: $\frac{\sum TP}{\sum TP+FP}$

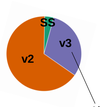
3. Single-cell RNA seq. dataset summary: Forebrain

A well-defined taxonomy is needed to successfully train a machine learning model. Here, we used the Allen Institute's mouse forebrain dataset, as it has well-defined cluster annotations and is created with data from a variety of scRNA-seq platform modalities: SmartSeq (SS), 10Xv2 (v2), and 10Xv3 (v3).

Forebrain taxonomy:



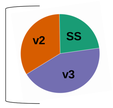
Full forebrain counts by modality:



Data Modality	Cells
SS	143,625
V2	2,422,766
V3	1,126,391

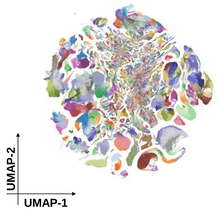
All v3 data were used for the training dataset.

Subsampled (test) forebrain counts by modality:



Data Modality	Cells
SS	143,625
V2	208,180
V3	270,545

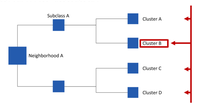
Subsampled data were used for the test datasets.



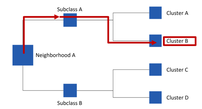
Forebrain dataset UMAP represented by 2156 clusters. Generated by Zizhen Yao.

4. Overview of flat and HKNN mapping algorithms

We need a K-Nearest Neighbor mapping algorithm (KNN) to classify unknown cells.
Algorithms tested:



"Flat" mapping
Calculates distance between test data and all training points at the cluster level. Ignores neighborhood/subclass labels.



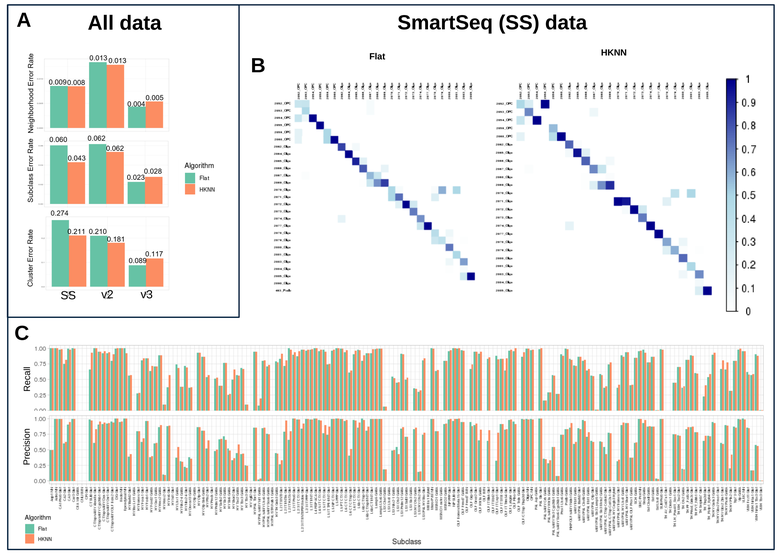
"HKNN" mapping
Begins distance calculations at the neighborhood level before evaluating at the subclass and cluster levels. Does not calculate distance metric between test data and all clusters.

Top.n.genes: the number genes' median expression data we look at to represent each grouping.
- This parameter can be optimized—the default is 15, shown in figure 5I.

Results

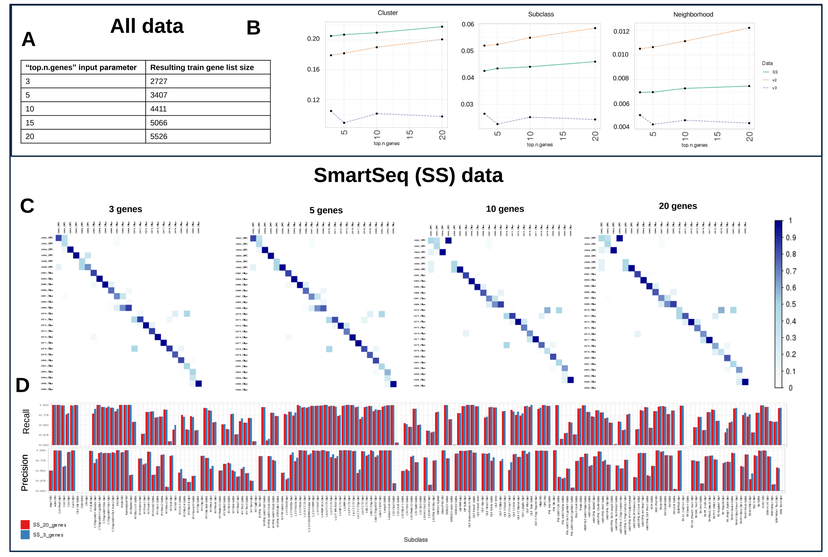
5. Optimize mapping algorithms

I. Comparison of algorithm error rates reveals HKNN mapping to be more accurate than flat mapping, except when using the v3 test dataset



(A) Neighborhood, subclass, and cluster level error rates for SS, v2, and v3 query datasets. (B) Cluster level confusion matrices for SS query dataset of the "Oligo NN" and "OPC NN" subclasses using flat and HKNN mapping algorithms. (C) Precision and recall rates at the subclass level on SS query data comparing flat (green) and HKNN (orange) mapping results.

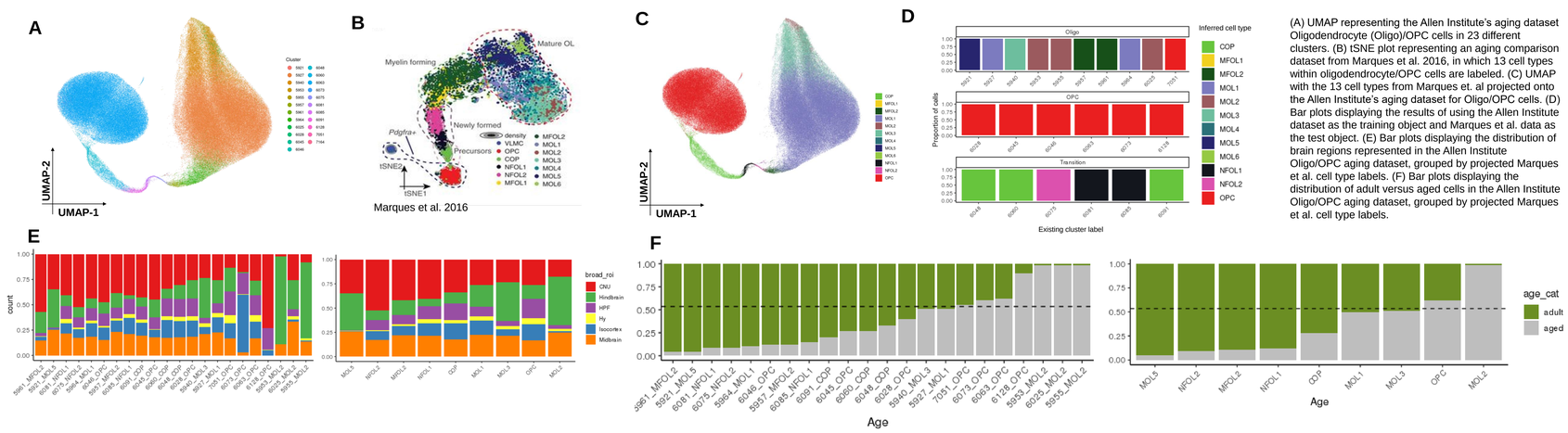
II. Smaller median gene expression lists (top.n.genes) in the training object produce more accurate mapping results



"top.n.genes" input parameter	Resulting train gene list size
3	2727
5	3407
10	4411
15	5066
20	5526

(A) Resulting length of train object median gene expression lists based on the top.n.genes input parameter. (B) Cluster, subclass, and neighborhood error rates for flat mapping with varying top.n.genes inputs on the SS, v2, and v3 query datasets. (C) Cluster level confusion matrices for SS query dataset of the "Oligo NN" and "OPC NN" subclasses when using a training object of 3, 5, 10, and 20 top.n.genes. (D) Precision and recall rates at the subclass level on the SS query data comparing flat mapping with 20 genes (red) to 3 genes (blue) as input.

6. Mapping application: External oligodendrocyte dataset characterizing adult and aged mouse brain cells



(A) UMAP representing the Allen Institute's aging dataset Oligodendrocyte (Oligo)/OPC cells in 23 different clusters. (B) tSNE plot representing an aging comparison dataset from Marques et al. 2016, in which 13 cell types within oligodendrocyte/OPC cells are labeled. (C) UMAP with the 13 cell types from Marques et. al projected onto the Allen Institute's aging dataset for Oligo/OPC cells. (D) Bar plots displaying the results of using the Allen Institute dataset as the training object and Marques et al. data as the test object. (E) Bar plots displaying the distribution of brain regions represented in the Allen Institute Oligo/OPC aging dataset, grouped by projected Marques et al. cell type labels. (F) Bar plots displaying the distribution of adult versus aged cells in the Allen Institute Oligo/OPC aging dataset, grouped by projected Marques et al. cell type labels.

7. Conclusions

- HKNN mapping tends to predict cell labels with greater accuracy than flat mapping, except when using a v3 train and v3 test dataset.
- Using lower numbers of genes to represent each grouping (top.n.genes) in the train object tends to lead to greater accuracy in predicting cell labels
- Using mapping algorithms on existing datasets can help define finer cell type groupings
- In Oligo/OPC cells, there is a noticeable enrichment of aged cells mapping to the MOL2 mature oligodendrocyte group, the majority of these cells derived from the hindbrain.
- We aim to determine the ideal number of genes for HKNN mapping in the future and use the most efficient algorithm to continue mapping external datasets to the Allen Institute's non-neuronal aging dataset. This work will aid in identifying the rare groupings within the non-neuronal aging dataset.

References

Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Mendanha Falcão, A., Xiao, L., Li, H., Häring, M., Hochgerner, H., Romanov, R. A., Gyllborg, D., Muñoz Manchado, A., La Manno, G., Lönnerberg, P., Floriddia, E. M., Rezayee, F., Ernfors, P., Arenas, E., Hjerling-Leffler, J., Harkany, T., … Castelo-Branco, G. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science (New York, N.Y.), 352(6291), 1326–1329. https://doi.org/10.1126/science.aaf6463

Acknowledgements

Thank you to Kelly, CK, Rohan, and all the Allen Institute employees supporting the intern program.

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG066027. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.