Deconvolving Rare Cell States from Spatial Transcriptomic Data

BMEN 4480 Final Project Report
Team 7: Anish Puligilla, Cody Slater, Joyce Zhou

ABSTRACT

The advent of spatial transcriptomics platforms have promoted new research endeavors seeking to understand how gene expression in cells varies with respect to the environment in the native tissue. Algorithms such as Robust Cell-Type Deconvolution (RCTD) have paved the way for scientists to deconvolve mixtures of cell types in spatial data and characterize the gene expression of specific cells within a mixture. However, algorithms such as this fail to account for the variety of states cell types can manifest as depending on other factors such as cellular environment. Microglia, due to their low tissue volume and relatively low abundance compared to other cell types in the central nervous system, have proven difficult to study. Here, to improve the characterization of microglial cell states, we use data collected from lipopolysaccharide (LPS)-injected mice to define cell state-specific clusters, integrate it with annotated clusters, and feed it into the RCTD algorithm applied to a spatial data from LPS-injected mice. Spatial mapping of microglia cell states using these extra clusters resulted in new information compared to baseline microglia detection, highlighting our approach's potential to reveal previously masked biological insights.

INTRODUCTION AND DATASETS

Spatial transcriptomic platforms such as 10x Visium have enabled researchers to study how gene expression in cells varies in the morphological context of their native tissue. In parallel, algorithms like Robust Cell-Type Decomposition (RCTD) have been developed to detect and deconvolve cell type mixtures in spatial data where gene expression measurements may contain contributions from multiple cells [1]. To do this, RCTD utilizes a single-cell RNA-seq (scRNA-seq) reference to define cell-type-specific profiles followed by a supervised learning approach.

In many cases, cells may adopt a distinctive phenotype on a spectrum of possible states. For example, depending on the environment and presence of immune-triggering stimuli, microglia can be homeostatic, activated, or partially activated. Microglia, specifically, have proven difficult to study given their overall small tissue volume and sparsity compared to other cell types in the central nervous system (CNS) [2]. Characterizing how microglial response varies spatially is imperative to our understanding of neuroinflammatory disease.

Here, we aim to improve characterization of microglial states in the brain using the techniques outlined above. We use flow sorted microglial data from Sousa et al collected from lipopolysaccharide (LPS)-injected mice to define separate clusters for microglia in different activation states [3] and integrate these clusters with an annotated transcriptomic atlas from the primary motor cortex in mouse generated by Yao et al [4], and subsequently feed this reference into RCTD applied to Visium data from brains of LPS-injected mice collected by Hasel et al [5]. A variety of open source labeled cell type datasets were tried, to serve as the gene library for deconvolution, including the Allen Institute Cell Type Database [6] and the original dataset used to demonstrate the functionality of the RCTD algorithm [1].

METHODS AND ALGORITHMS

Defining Microglial Clusters

A. Preprocessing. Raw count data for control and LPS treatment groups were loaded into anndata

objects and then concatenated into a single anndata object. Cells with less than 1000 counts and cells with mitochondrial counts greater than 15% were filtered out. Genes not expressed in at least 1 cell were also removed. Counts were scaled to counts per million (CPM) and log transformed. Library size effects were regressed out.

B. Dimensionality reduction, tSNE, and clustering. Principal Component Analysis (PCA) on the dataset was performed. After inspecting the PCA variance ratios, the first 4 principal components (PCs) were subsequently used for tSNE. Two clustering methods were tested, k-means and leiden. For k-means clustering, k = 2, 3 were tested. For leiden clustering, nearest neighbors was set to 15 and resolution to 0.3. Leiden clusters were used for subsequent downstream analysis.

C. Differential expression analysis. The top five differentially expressed genes (DEGs) in each cluster were found using Wilcoxon rank-sum. Cell clusters were subsequently labeled as homeostatic microglia, activated microglia, or intermediate activated microglia.

D. Diffusion pseudotime analysis. Data was annotated with a root cell (first cell in control group), and diffusion pseudotime analysis was performed using 10 diffusion components and plotted on the first 2 diffusion components.

E. Export files in preparation for RCTD. The raw counts matrix for cells kept after preprocessing was converted to a dataframe and exported to a csv file. Microglial activation cluster labels were exported to a metadata csv file. Total number of unique molecular identifiers (UMIs) were calculated using the sum of the gene counts after the processing steps above.

F. Integration with annotated motor cortex mouse dataset. Raw annotated count data for the mouse motor cortex data was loaded into an anndata object. Cells labeled with subclass 'Micro-PVM' were removed. This dataset was then merged with the newly annotated microglia data into a single anndata object using concatenation. The merged dataset was log normalized. Highly variable genes (HVGs) were used for feature selection. Batch effects and library size effects were regressed out. Data was visualized in UMAP using manhattan distance, with n_neighbors set to 30 and min_dist to 1.

Baseline Analysis of Spatial Data

A. Generation of UMI Counts. For each of the six visium slices, three LPS treated and three saline, the total UMI counts for each pixel were generated using basic Seurat spatial transcriptomic visualizations.

B. Generation of Spatial Transcriptomic Metadata. A separate spatial metadata file was then generated that stored the number of UMIs for each pixel, labeled with a unique barcode. A separate file with the x and y coordinate locations for each pixel was also generated for each slide.

C. Visualization of top Differentially Expressed Genes on Spatial Images. For each of the saline and LPS treated visium slices, the top 5 DEGs, identified as part of the microglial cluster identification process, were visualized using a low resolution tissue image. For the second LPS treated slice, no low resolution image was available in the dataset and the high resolution image was used in its place.

Applying Microglial Clusters to Deconvolve Spatial Data Using RCTD

A. Preprocessing and quality checks of reference data. Using the data files created above, each file was manually checked for a series of quality control metrics, including verification of barcode names across files and deletion of duplicate/corrupted gene data from the original dataset. For the single cell reference data, three CSV files were generated for import into the RCTD framework: 1) a count matrix for all gene counts for each cell included as a reference. If multiple datasets were combined together to create the reference, the gene order was aligned with the primary dataset and the additional cells were appended as columns on the count matrix. The counts were untransformed, raw count data. 2) A cell type dictionary was generated in which each uniquely barcoded cell in the reference count matrix was labeled with the known cell type. 3) A UMI count for the total number of post-filtered gene counts in each uniquely barcoded cell in the reference. For the final implementation of this demo algorithm, a truncated, labeled dataset provided by the authors

was used. This dataset was chosen because of the feasibility of running the deconvolution algorithm and the ability to iterate the results in a more controlled manner. The labeled dataset included ~25 cells of several common brain cell types, including neurons, astrocytes, oligodendrocytes, polydendrocytes, and a generic microglia-macrophage combined cluster. A total of ~1200 labeled microglia from the previous analysis were included in the reference dataset. The mean gene expression and variance for each microglia class was calculated by the RCTD algorithm.

B. Preprocessing and quality checks of the spatial data of interest. For the Visium 10x dataset used in this analysis, two CSV files were generated for import into the RCTD algorithm: 1) a file containing the x and y coordinates for each of the unique pixels in the dataset. 2) a file containing the total number of gene counts per pixel. UMI counts were obtained by summing the gene expression matrix.

IMPLEMENTATION

Defining Microglial Clusters

Microglial dataset analysis was implemented in Python primarily using scanpy, with k-means clustering performed using sklearn. The Seurat object containing the mouse primary motor cortex dataset was converted to an h5ad file using sceasy in R, before loading into Python.

Baseline Analysis of Spatial Data

Generation of UMI counts and differential gene expression tables was done in R using RStudio using the Seurat objects created from the matrix, barcode, and feature data. Visualization of differentially expressed genes was also performed using R in RStudio with the Seurat package using functions SCTransform() and SpatialFeaturePlot().

Applying Microglial Clusters to Deconvolve Spatial Data Using RCTD

The coordinates, counts, and UMIs were loaded into a puck object using the SpatialRNA() function in R, using RStudio. Subsequent visualization of the pixel expression data was accessed through this object. The count matrix, cell type dictionary and UMI counts for the referenced single cell data were loaded into a reference object using the R function Reference().

Due to computational intensity and time constraints, the original count matrices with ~30,000 genes were gradually reduced until we arrived at the approximately ~480 genes used in the simplified dataset used by the original RCTD authors.

Once the data met the quality checks discussed above, the reference and puck objects were passed into the RCTD algorithm, using the "full" analysis mode, in which no predetermined number of cells per pixel is set. For the highly simplified dataset, application of the algorithm to a single Visium slide took approximately ~10 minutes to run on a dedicated desktop with 32 cores, 128gb RAM, and a 3080ti GPU.

RESULTS

Microglial Activation State Profiling

Following preprocessing and PCA (Figs. 1A, B), microglia cells can be seen to partition across tSNE space by treatment (Fig. 1C).

Both k-means and leiden clustering generated reasonable clusters based on visual inspection of the tSNE embedding (Fig. 2). Three clusters can clearly be seen, in agreement with results from Sousa et al, and indeed, three clusters were identified using k = 3 for k-means or using leiden.

Because the leiden algorithm generated robust clusters and is generally advantageous over k-means for not needing prior assumptions about k or having sensitivity to outliers, we moved forward with leiden clusters for the remaining downstream analysis.

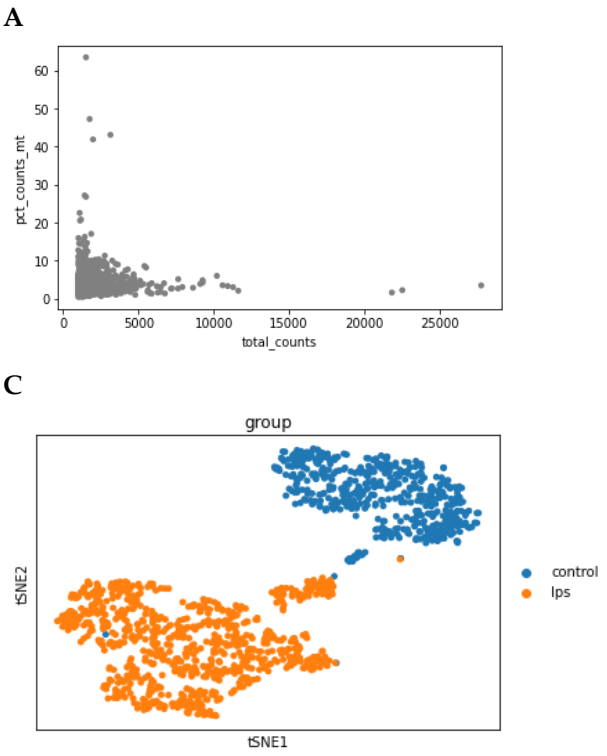

Fig. 1. Microglia data preprocessing and visualization. (A) Percent mitochondrial counts vs total counts (library size). (B) PCA variance ratio plot for top 30 PCs, log scaled. (C) tSNE embedding colored by treatment (control or lps).

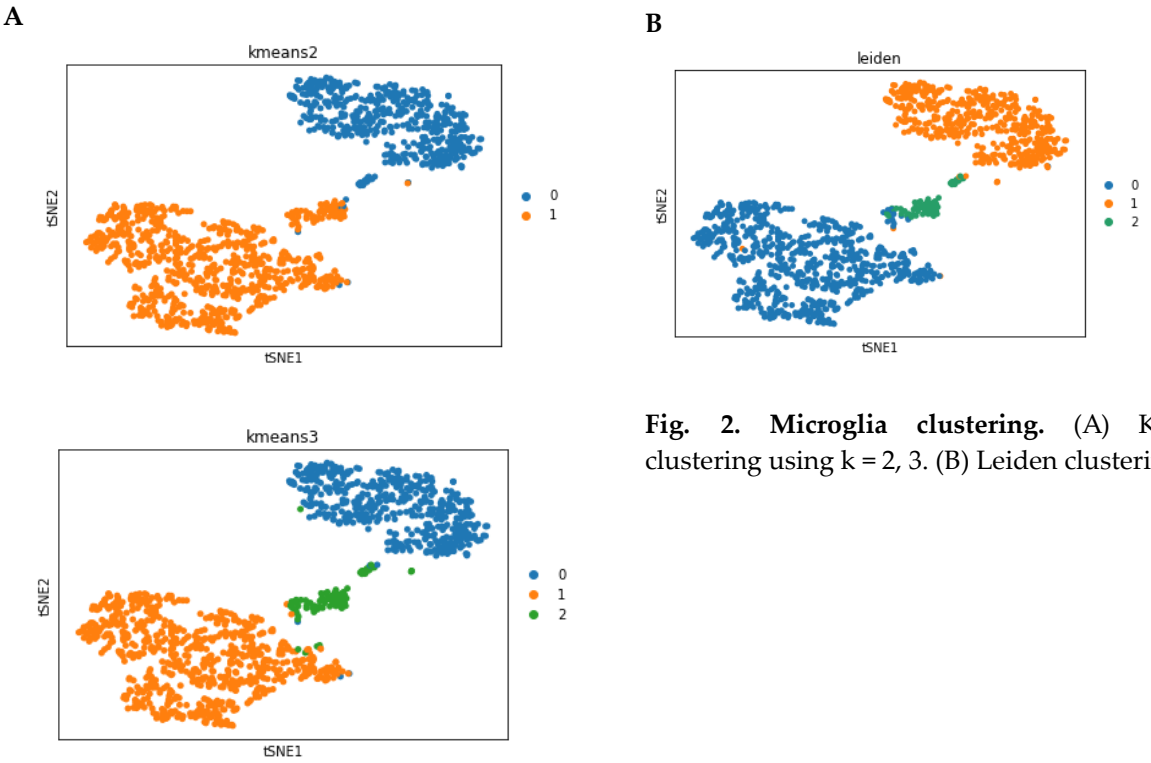

Fig. 2. Microglia clustering. (A) K-means clustering using k = 2, 3. (B) Leiden clustering.

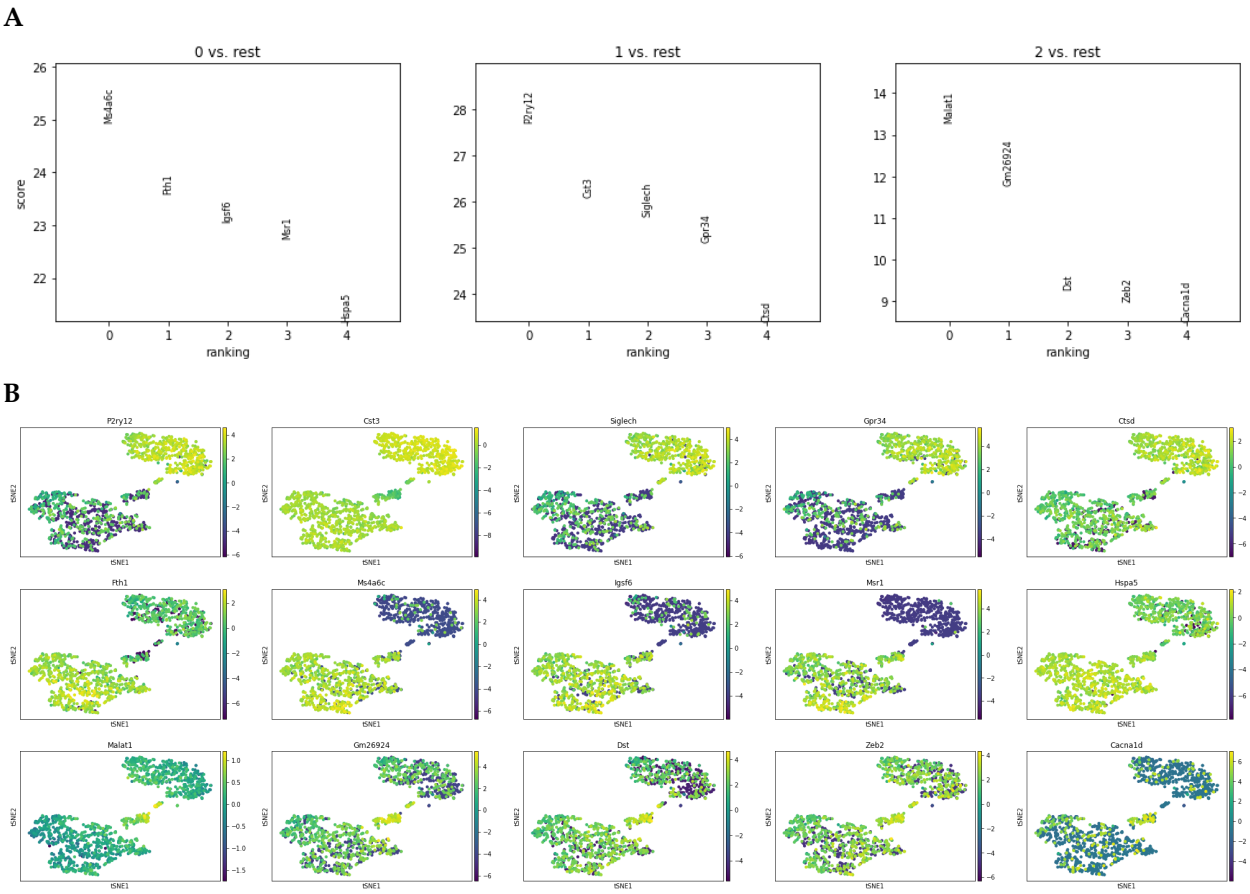

Fig. 3. Microglia DEGs. (A) Top 5 DEGs in each cluster versus all other clusters. (B) tSNE plots colored by DEG expression. The top row are DEGs in cluster 1, the middle row are DEGs in cluster 0, and the bottom row are DEGs in cluster 2.

The top differentially expressed genes (DEGs) within each of the three clusters were identified (Fig. 3A) and their expression levels across all cells were visualized in tSNE space, showing upregulation of each cluster's DEGs within each respective cluster (Fig. 3B).

Cell clusters were subsequently labeled as homeostatic microglia, activated microglia, or intermediate activated microglia based on Sousa et al (Fig. 4A), and we see that mice exposed to LPS exhibited a down regulation in homeostatic function and an upregulation in inflammatory markers (Fig. 4B).

Diffusion pseudotime analysis revealed that the activation response of the microglia can be understood as following a trajectory from the resting state to the activated, immuno-responsive state (Fig. 4C). Examining the pseudotime plot colored by microglial clusters, we see the homeostatic state transition into the activated state followed by the intermediate activated state, reflecting that the microglial cells in the intermediate activated state have a comparatively delayed response.

Lastly, UMAP plots in Fig. 5 show all annotated cell types from the primary mouse motor cortex dataset before (Fig. 5A) and after (Fig. 5B) integration of the newly annotated microglial data.

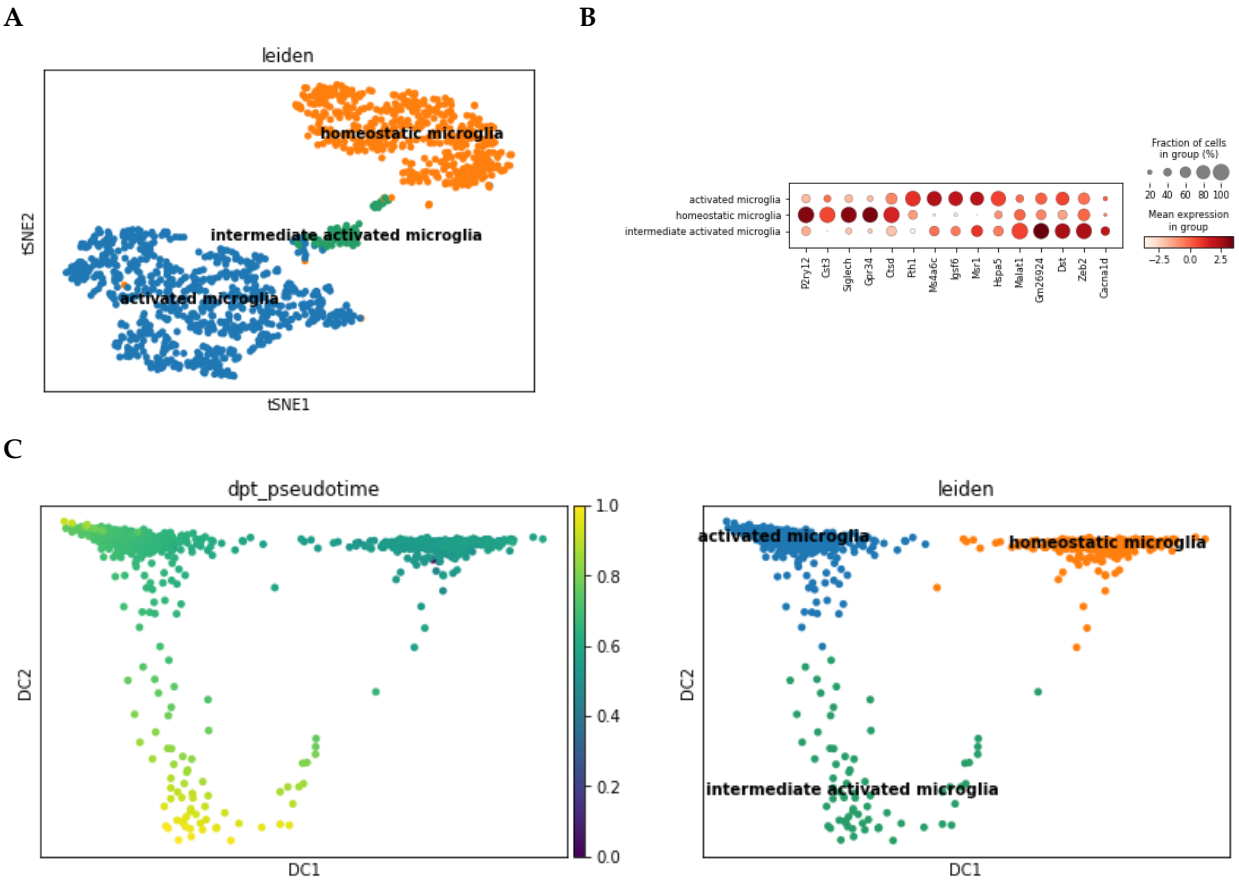

Fig. 4. Labeling microglial activation states. (A) Leiden clusters relabeled by microglial activation state. (B) Dotpot visualization of top 5 DEGs for each state. (C) Diffusion pseudotime plots.

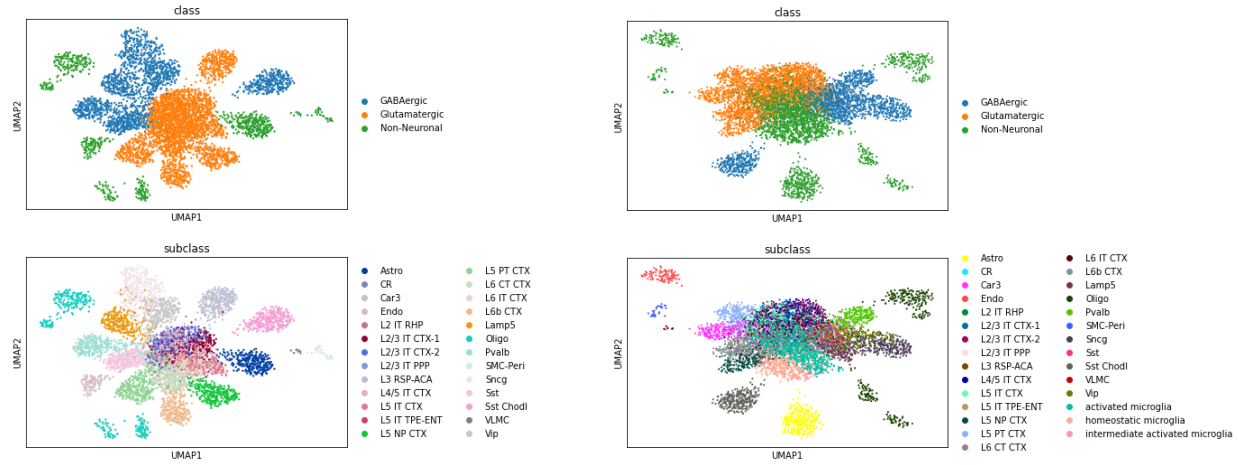

Fig. 5. Integrating microglia clusters with other cell type clusters. (A) Mouse motor cortex cell type clusters (after removal of microglia-labeled cells) visualized in UMAP before combining with new microglia clusters. (B) Cell type clusters visualized after integrating new microglia clusters.

Seurat Exploration of Microglial Markers

Following microglial cell cluster identification and identification of the top 5 differentially expressed genes in each cluster, representing different activation states of microglial cells, gene expression was spatially visualized utilizing the acquired low resolution tissue slice image. In the activated microglial cell cluster, the genes Fth1 and Hspa5 appear to have widespread expression with the remaining two genes only expressed in certain pixels. In the intermediate activated cluster, most of the differentially expressed genes appear to be expressed throughout the tissue in varying levels. Among the genes most expressed in the homeostatic microglial cell cluster, only two, Cst3 and Ctsd were expressed widely in the tissue sample. The remaining three genes were only expressed in a smattering of pixels throughout the slice.

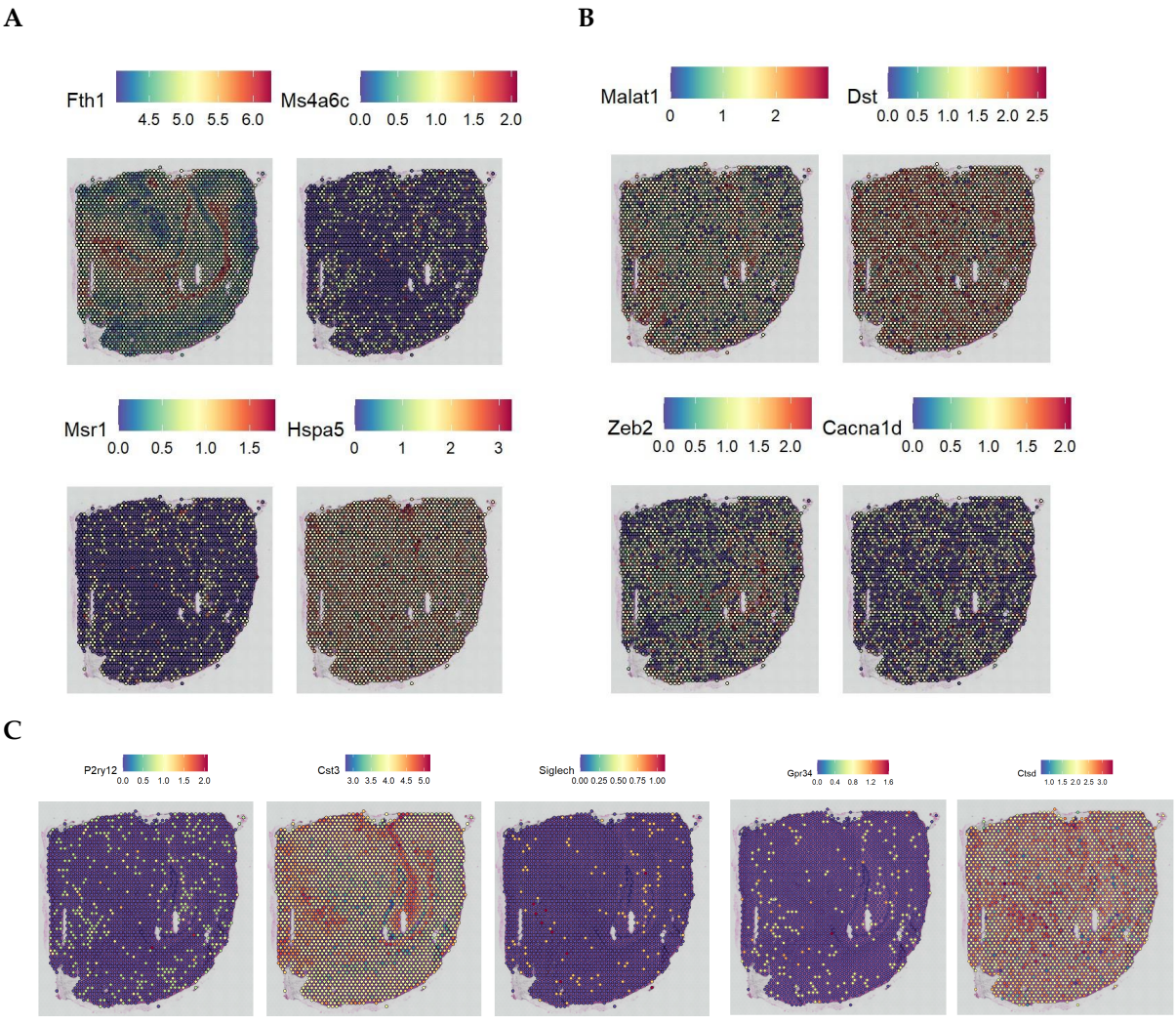

Fig. 6. Spatially visualizing the five most common differentially expressed genes (DEG) in each of the three clusters of microglial cells (Fig. 3) in an LPS treated visium slice. (A) Top 4 DEGs in the activated microglial cell cluster. (B) Top 4 DEGs in the intermediate activated microglial cell cluster. (C) Top 5 DEGs in the homeostatic microglial cell cluster. Similar images for other LPS and Saline treated visium slices are available online along with our code.

Deconvolution of Spatial Data with Generically Labeled Reference Data

In order to establish a baseline for comparison of subsequent algorithms, the single cell reference data containing only a generic microglia-macrophage group (in addition to the other generic cell type definitions) was used with the RCTD tool to estimate the weights (fractions) of total UMIs per pixel contributed by different cell types in the reference dictionary. Fig. 7 shows the results.

Deconvolution of Spatial Data with Newly Defined Microglial Clusters

Following baseline characterization of the ability of the generic clusters to identify microglia in the control and treatment datasets, the custom derived microglial state definitions identified previously were included in the deconvolution algorithm and the three separate microglia clusters were identified. The performance in the control and LPS-treated datasets is shown in Fig. 8. Each microglial cluster (homeostatic, intermediate activated, or activated) was identified independently through maximum likelihood estimation.

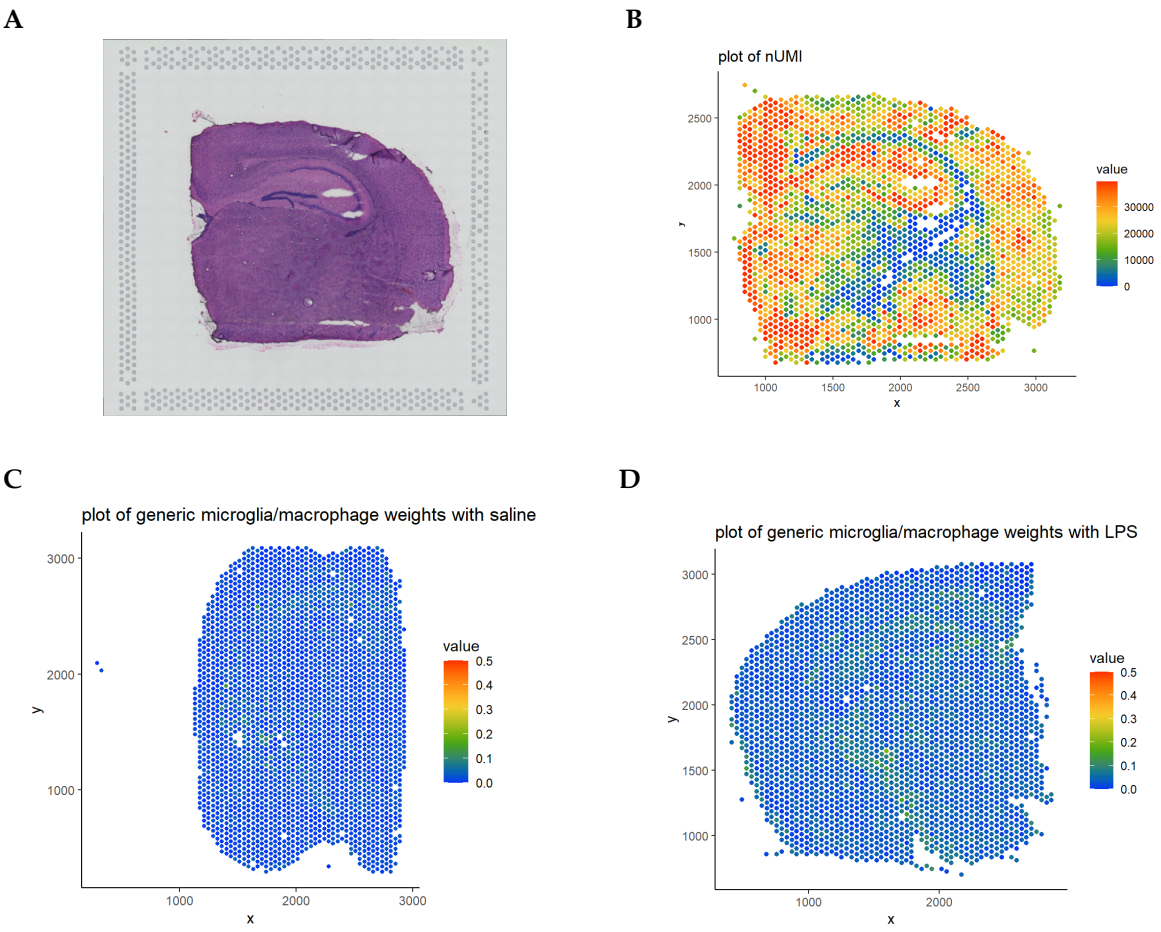

Fig. 7. Application of the RCTD algorithm to sample Visium data using generic and custom cluster labeling. (A) H&E stain of a sample Visium slide. (B) Number of UMIs per pixel, as visualized through the spatial data loaded into the puck object. (C) Sample results from the Visium brain slice of a mouse treated with saline using the generic clusters to perform deconvolution (D) Sample results from the Visium brain slice of a mouse treated with LPS using the generic clusters to perform deconvolution; each pixel color demonstrates the estimated proportion of the specified cluster type.

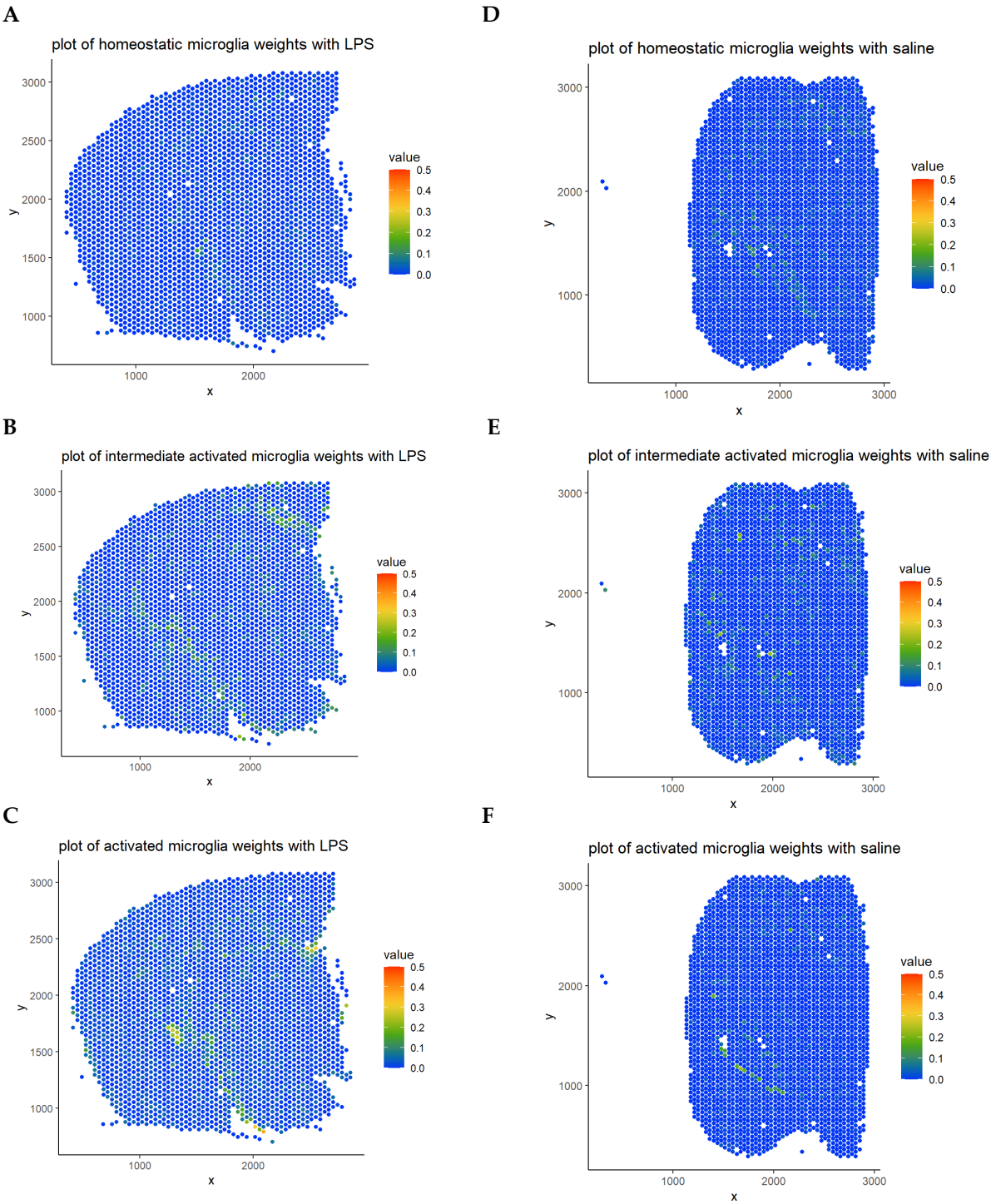

Fig. 8. Application of the RCTD algorithm to sample Visium data using new microglia cluster labeling. (A-C) Pixel-level weighting of estimated homeostatic, intermediate-activated and activated microglia in an LPS peripherally injected mouse brain. (D-F) Pixel-level weighting of estimated homeostatic, intermediate-activated, and activated microglia in a saline controlled mouse brain.

DISCUSSION

While deconvolution methods for resolving cell mixtures per pixel in spatial data have been developed, these methods in some cases can be subject to bias in identifying clusters of cell types, and in particular they can be susceptible to losing granularity and potentially masking more rare cells that may exist in lowly abundant heterogeneous states. In addition, this approach does not take full use of the immense amount of publicly available genomic datasets in which cell states of interest can be profiled with much higher resolution and retroactively applied to techniques such as Visium spatial transcriptomics which captures transcriptomic information at a lower resolution, but could benefit from an a priori understanding of cell states of interest.

Microglia in the activated immune-responsive state, for instance, may exist only in a rare subset of cells close to a hypothetical region or pathology of interest in the brain. The specific activated phenotype of an individual microglial cell may also vary depending on proximity to the pathological site of interest. Therefore it may be useful to resolve spatial transcriptomic data using a spectrum of microglial clusters undergoing varying degrees of activation to understand changes in cell state in the context of spatial location and morphological disease features of interest within the brain. This informed the aim of our project, which was to see whether defining the activated, rare cell state of interest prior to applying a deconvolution algorithm would allow us to derive new insights on microglial activation spatially.

A study by Chen et al also sought to understand changes in gene expression in the vicinity of pathogenic features, using spatial transcriptomics combined with *in situ* sequencing on mouse and human brain to look at the transcriptional changes characterizing the inflammatory response in small tissue domains around amyloid plaques of Alzheimer's disease [7]. Whereas the authors take the more classical approach of performing stand-alone clustering analysis on their data, the novelty of our approach here lies in utilizing available genomics data to create and define cell states of interest we know should exist. We emphasize that this may be especially useful for more rare cell states that are susceptible to getting lost in spatial data. The task then becomes to match cells from a given spatial transcriptomic dataset against a defined range of universal clusters existing along a continuum of cell states, not just discrete cell types.

Our attempt here to create this algorithmic approach has yielded interesting results. When compared to the generically clustered reference dataset, the custom labeled microglial state clusters do appear to highlight subpopulations of microglia across the spatial data. The LPS dataset deconvolved with the generic labels in Fig. 7D does identify many of the same spatial regions as including microglial cells as the custom labels, but the results of the latter resulted in highlighted heterogeneous subpopulations of homeostatic, intermediate-activated, or activated microglia, see Fig 8A-C. Similar results were seen for other Visium datasets and the results can be seen in the online code repository.

For the saline control data, which should be expected to have a primary population of homeostatic microglia, the deconvolution algorithm with the custom clusters seemed to perform suboptimally, perhaps indicating the homeostatic labeled microglia were not generalizable to our spatial dataset.

Limitations

There are a number of important limitations that need to be addressed when interpreting the results presented here. First, the resource intensive nature of this analysis forced us to gradually reduce the number of included genes in our reference library until we were only left with <500 genes, which were dictated by the limitations of our generically labeled dataset, which did not have some of the top upregulated/downregulated genes of interest identified in previous analysis. It is possible that a more targeted selection of genes in the reference library would further improve the performance of this approach.

Another major limitation is the fact our generic labels lack robustness or specificity for our data of interest. One of the datasets we attempted to use included much higher resolution per cell type and included mean gene expression data for >300 different cell types (only one of which was dedicated to microglia). The inclusion of a large dataset like this would add costly penalties in terms of computational time, but may improve deconvolution performance and significantly improve the ability to visualize diverse cell types across the entirety of spatially collected genomic data.

Future Work

Our results here demonstrate that this framework can be used for mapping finer-resolution cell states onto spatial data and uncovering previously masked activation cell-state specific spatial localization. For this project, we used data collected from a study investigating the response of microglia to LPS, a pro-inflammatory stimulus. A more interesting application of our method would be in pathogenic slices, from brain tumor or neurodegenerative disease samples for example. This may generate further insights into how the inflammatory response varies in the vicinity of diseased regions.

Our approach also has broad applicability to other cell types that may also adopt different phenotypes along a spectrum. For example, epithelial to mesenchymal transition (EMT) is a process by which epithelial cells transform into mesenchymal cells. This is a dynamic and reversible process, and cells may be in intermediate stages of this transition to varying degrees depending on environmental cues as well as individual cell heterogeneity [8]. Therefore, it may be interesting to spatially visualize cells in early, intermediate, and late stage EMT in biological processes such as embryonic development or tissue regeneration.

CONTRIBUTIONS

Joyce performed the analysis on the scRNA-seq microglia data to characterize microglial states. Anish visualized the spatial data from the Visium dataset and generated the baseline results for comparative analysis. Cody integrated the microglia cluster analysis with the spatial data and performed deconvolution analysis.

CODE

All code for this project is available at https://github.com/joycekzhou/BMEN4480_FinalProject.git.

REFERENCES

[1] Cable, D.M., Murray, E., Zou, L.S. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* (2021).

[2] Sankowski, R., Monaco, G., Prinz, M. Evaluating microglial phenotypes using single-cell technologies. *Trends Neurosci* (2022).

[3] Sousa, C., Golebiewska, A., Poovathingal, S.K. *et al.* Single-cell transcriptomics reveals distinct inflammation-induced microglia signatures. *EMBO Reports* (2018).

[4] Yao, Z., Liu, H., Xie, F. *et al.* An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv* (2020). Seurat object downloaded from from https://azimuth.hubmapconsortium.org/references/#Mouse%20-%20Motor%20Cortex

[5] Hasel, P., Rose I.V.L., Sadick, J.S., Kim. R.D. *et al.* Neuroinflammatory astrocyte subtypes in the mouse brain. *Nat Neurosci* (2021).

[6] Z. Yao, C. T. J. van Velthoven, T. N. Nguyen, J. Goldy, A. E. Sedeno-Cortes, F. Baftizadeh, D. Bertagnolli, T. Casper, M. Chiang, K. Crichton, S. L. Ding, O. Fong, E. Garren, A. Glandon, N. W. Gouwens, J. Gray, L. T. Graybuck, M. J. Hawrylycz, D. Hirschstein, M. Kroll, K. Lathia, C. Lee, B. Levi, D. McMillen, S. Mok, T. Pham, Q. Ren, C. Rimorin, N. Shapovalova, J. Sulc, S. M. Sunkin, M. Tieu, A. Torkelson, H. Tung, K. Ward, N. Dee, K. A. Smith, B. Tasic and H. Zeng, Cell 2021, 184, 3222-3241 e3226.

[7] Chen, W., Lu, A., Croissants, K., *et al.* Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell* (2020).

[8] Pal, A., Barrett, T.F., Paolini, R. *et al.* Partial EMT in head and neck cancer biology: a spectrum instead of a switch. *Oncogene* 40, 5049–5065 (2021). https://doi.org/10.1038/s41388-021-01868-5