

Brain Cell Type Specific Gene Expression and Co-expression Network Architectures

Andrew T. McKenzie, Minghui Wang, Mads E. Hauberg, John F. Fullard, Alexey Kozlenkov, Alexandra Keenan, Yasmin L. Hurd, Stella Dracheva, Patrizia Casaccia, Panos Roussos, and Bin Zhang

Supplementary Methods

Read mapping and data pre-processing

For each human or mouse RNA-seq datasets analyzed, we downloaded published raw sequencing data from GEO (**Table 1**). While either normalized gene expression measures or gene-level read counts data were provided along with the original publications, the RNA-seq data from the same species were processed by different pipelines in different studies, thus resulting in different gene expression quantification metrics, rendering complexity and incompatibility in conducting a cross-dataset comparison. To simplify the downstream data analysis, we reprocessed the sequencing alignment and gene expression normalization using a unified and efficient pipeline built upon the STAR aligner¹, featureCounts², and R/Bioconductor package edgeR³. With this pipeline, the raw sequencing reads were aligned to either human hg38 genome or mouse mm10 genome using the STAR aligner. For human datasets, GENCODE gene annotation model GRCh38.p2 release 22 was used to assist with the alignment of reads to known gene features. For mouse datasets, the UCSC mm10 gene annotation model was used. Following read alignment, featureCounts was used to quantify the gene expression at both the gene and exon level. For the single cell data sets, the scater R package (version 1.0.4)⁴ was used to calculate the total number of features identified in each sample, which was used as a covariate in the differential expression models, as it may reflect biases in the initial amount of RNA prior to amplification. For the Darmanis *et al.* data set, 3 outlier samples were detected and removed on the basis of principal component analysis using the scater R package (SRR1974559, SRR1974625, and SRR1974648). The gene level read counts data was normalized using the quantile approach^{3,5}, with a percentile of 0.75 for the cell population data sets and 0.99 for the single cell data sets, to adjust for sequencing library size difference.

Gene symbol conversion

For all three murine data sets, the gene symbols were converted from MGI symbols to HGNC symbols by using the Ensembl database, accessed through biomaRt (version 2.28)⁶. If no homolog existed, then the original mouse symbol was retained. In the case of multiple homologous genes, we selected the human gene with the highest homology percentage based on protein coding region DNA divergence. In the case that multiple transcripts mapped to the same gene symbol, the gene symbol with the maximum average expression count was retained for subsequent analyses⁷.

Comparison of cell type signatures to PubMed text mining results

For the top 100 most enriched genes in each cell type, we performed PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>) searches and counted the number of results for the search “Gene AND CellType” (**Supplementary File 2**), where Gene refers to the character string of the enriched gene symbol and CellType refers to the character string of the enriched cell type, either “astrocyte”, “endothelial”, “microglia”, “neuron”, “oligodendrocyte”, or “oligodendrocyte precursor”. We used the Mann-Whitney U test to compare the number of PubMed mentions found among the corresponding most enriched gene set, compared to the genes most enriched in all of the other cell types, with the exception of oligodendrocytes and OPCs, which were not included in each other’s control sets for comparison because of their similarity. We also found the Spearman correlation between the enrichment for each gene in that cell type and the number of PubMed mentions for the gene symbol/cell type combination for each cell type.

Comparison of the three cell-type associated gene expression measurements

We first selected genes such that they needed to be within the top 1000 genes as ranked by one of the measures at a time. We used this filtering method because, on one hand, we needed to remove low-ranked genes because the measures were not designed to be sensitive to differences in gene ranks at the lower ranks, because the expression of a gene in a given cell type for all data sets may be negligible. On the other hand, if we were to filter genes such that they ranked highly in at least one of the categories, we would have introduced a spurious association between the two variables, as a result of Berkson's bias. Therefore, we calculated the Spearman

(rank) correlation between gene rankings for each of the three measures after filtering for the top ranked genes in one measure at a time, for a total of unique six comparisons, for each of the six cell types.

Estimation of relative cell type proportions from bulk RNA expression data

To make relative cell type proportion estimates, we adapted the previously validated singular value decomposition (SVD) method from CellCODE⁸, which has been implemented in the BRETIGEA (BRain cEll Type specIfic Gene Expression Analysis) R package (version 1.0). Specifically, after scaling the data, the algorithm in BRETIGEA calculates the first singular vector of a variable number of cell type-specific (marker) genes as an estimate of the relative cell type proportion in those samples. Because the signs of this singular vector can be reversed, we next find the mean correlation between the singular vector and each of the marker genes, and if this is less than zero, then the sign of all of the values in the singular vector are switched. Note that these estimated values can be less than zero for a given sample, which is part of why it is a relative cell type proportion estimate, rather than an absolute cell type proportion estimate. BRETIGEA also offers principal component analysis (PCA) as an option for cell type relative proportion estimation, along with sets of 1000 top consensus cell-specific genes for each cell type.

We used matched RNA expression and immunohistochemistry marker data from the Aging, Dementia, and TBI study from the Allen Brain Atlas⁹ as an additional validation of the marker genes identified in bulk brain gene expression data. In this study, investigators performed both RNA-seq and immunohistochemistry (IHC) quantification for protein levels of the astrocyte marker gene GFAP and the microglia marker gene IBA1 in brain samples from the same donor across four brain regions (frontal white matter, hippocampus, temporal cortex, and parietal cortex). We downloaded the normalized RPKM values from the study website (<http://aging.brain-map.org/download/index>), which has been adjusted for RNA integrity number (RIN) and batch. The IHC quantifications provide an independent measurement for the relative levels of the cell types in each brain sample, which can be used to validate our relative cell type proportion estimates from the RNA expression data. Using all four brain regions available in the Allen Brain Atlas data set, we first calculated 100 SVD-based estimates of relative cell type proportions for astrocytes and microglia in each sample using the top

1-100 marker genes. We correlated all of these estimates with the independent IHC quantifications for GFAP and IBA in those samples. We also found the rank correlation between each of the individual top 100 marker genes for astrocytes and microglia with the independent IHC quantifications for GFAP and IBA1, respectively. As an additional exploratory analysis, we estimated the relative proportion of astrocytes and microglia using the SVD-based method *within* each of the 4 individual brain regions and calculated the rank correlations of these estimates with the IHC quantifications for GFAP and IBA1, respectively, in each of those brain regions.

ATAC-seq validation of novel cell markers

To validate our novel cell marker genes, we performed ATAC-seq on 50mg of frozen postmortem brain tissue from the dorsolateral prefrontal cortex from four controls (i.e. individuals that had no history of neuropsychiatric disease or substance abuse). Human brains from normal control Caucasian subjects without head trauma were collected at autopsy within 24 h after death at the Department of Forensic Medicine, Semmelweis University and at the National Institute of Forensic Medicine (Karolinska Institutet). All material was obtained under approved local ethical guidelines. Nuclei were processed for fluorescence activated nuclear sorting (FANS) as described previously¹⁰. ATAC-seq was performed on FANS sorted nuclei to assess chromatin accessibility following an established protocol¹¹. We mapped ATAC-seq reads with STAR 2.50 and the following settings: `--alignIntronMax 1 --outFilterMismatchNmax 100 --alignEndsType EndToEnd --outFilterScoreMinOverLread 0.3 --outFilterMatchNminOverLread 0.3`. To rule out sample mislabelling and contamination, genotypes were called with GATK 3.50 and the genetic concordance between all combinations of samples calculated. Peaks were subsequently called with the MACS 2.1 narrow peaks algorithm with the following settings `--keep-dup all --shift -100 --extsize 200 --nomodel`. We created a consensus set of peaks by keeping only those that were found in two or more samples. Read counts were then quantified within these peaks using RSubread 1.15 and the following settings: `allowMultiOverlap=F, isPairedEnd=T, strandSpecific=0, requireBothEndsMapped=F, minFragLength=0, maxFragLength=2000, checkFragLength=T, countMultiMappingReads=F, countChimericFragments=F`. This yielded a peak by sample expression matrix, from which peaks mapping to the ENCODE blacklisted regions were excluded. Peaks that had very few read

counts (defined as peaks that had less than 1 count per million in 90% or more of the samples) were then discarded. For the remaining peaks, we normalized the read counts using "TMM" normalization followed by conversion of read counts to counts per million. We then constructed a peak by gene mapping by considering a peak mapped to a gene if a TSS overlapped the peak. Most often a gene overlapped, at most, one peak, but when it overlapped multiple peaks due to multiple transcripts we took the sum of the reads within the peaks. Additionally, genes not having a TSS overlapping an ATAC-seq peak were left out of further analyses. This yielded a gene by sample matrix of read counts. Subsequent averaging across samples from the same cell type yielded a gene by cell type matrix of read counts. This matrix was finally normalized by the sum of read counts for the given gene, resulting in a matrix for each gene indicating how large a fraction of the ATAC-seq read counts originated from a given cell.

Multiscale network analysis in the single cell RNA expression data sets

MEGENA networks were constructed from genes expressed in at least half of the samples for each cell type in each data source by using the R package MEGENA (version 1.3.4-1). Briefly, Pearson correlation coefficients (PCCs) were first computed for all gene pairs. The gene pairs with absolute PCCs larger than 0.3 were ranked and iteratively tested for planarity to grow a Planar Filtered Network (PFN). Multiscale clustering analysis was conducted with the resulting PFN to identify coexpression modules at different network scale topologies using the default parameter settings of the package. We visualized the networks derived from the MEGENA analysis using the prefuse force directed layout option in Cytoscape (version 3.2.1).

To visualize the gene ontology (GO) enrichment of distinct modules in the multiscale network, we filtered for modules containing between 5 and 2000 gene members, and selected only those modules with at least one gene that was specific to that module, thus removing exclusively parent modules. We used the moduleGO function in DGCA¹² (version 1.0.1) to perform gene ontology (GO) enrichment analysis on gene modules, which leverages the GOstats (version 2.34)¹³ and org.Hs.eg.db GO annotation (version 3.1.2) R packages. To identify GO terms with potential brain cell type-specific activity, we filtered for those GO terms with less than 500 gene symbols. We adjusted the enrichment p-values for all GO terms in each module using

the Benjamini-Hochberg (BH) method. We used Fisher's Exact Test (FET) to identify cell type enrichments for each module, using the top 500 genes ranked by consensus cell type specificity from each cell type in the human data sets as the cell type enrichment signatures, and adjusted the module enrichment p-values for each cell type signature using the BH method. The background universe size used in the FETs was the number of unique gene symbols found in both the human and mouse data sets. To identify module conservation, we found the FET enrichment of the intersection of all the modules identified in each cell type from the Darmanis *et al.* human data set with all the modules identified in each cell type from the Tasic *et al.* and Zeisel *et al.* mouse data sets. We counted the number of significantly overlapping (BH-adjusted $p < 0.05$) modules in each of these comparisons. We corrected for the total number of overlapping modules identified within each mouse cell comparison set by normalizing the resulting overlap matrix by column sums.

Supplementary Data

Supplemental File 1. Cell type associated gene rankings across data sets.

The Excel file contains tab-separated tables with the rankings for all three cell type associated measures, i.e. cell type-enrichment, cell type-specificity, and cell type-expression, across all data sets, human-only, and mouse-only, in all 6 cell types.

Supplemental File 2. PubMed text mining results.

The Excel file contains tab-separated tables with the top 100 most enriched genes in each of the six cell types, as well as the number of PubMed abstracts for a search of each gene symbol and the cell type in combination, as well as the gene symbol alone.

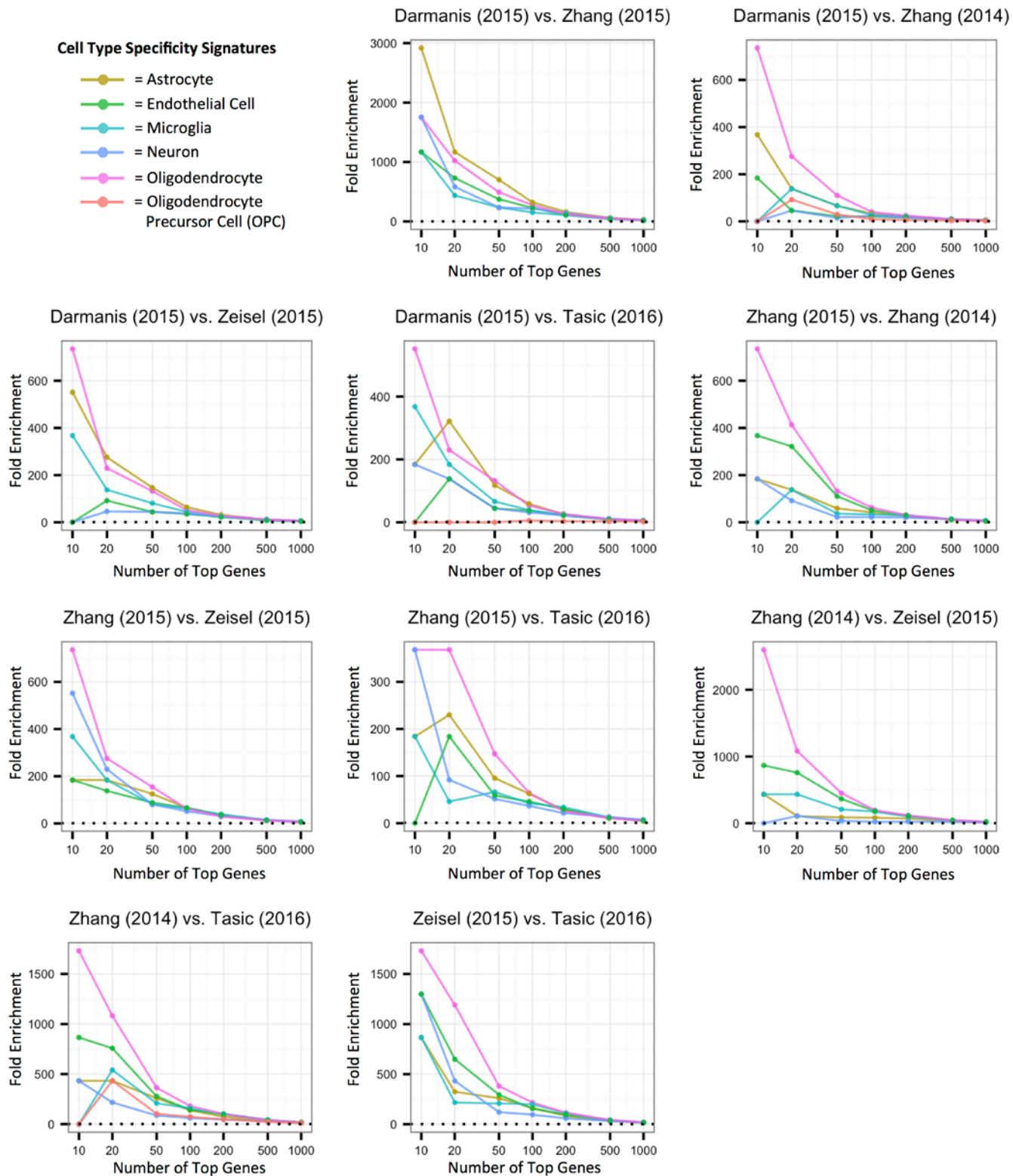
Supplemental File 3. MGI phenotypes enriched in the cell type specific marker genes.

The Excel file contains the statistics from the enrichment test of MGI phenotype gene signatures in the cell type specific marker gene signatures.

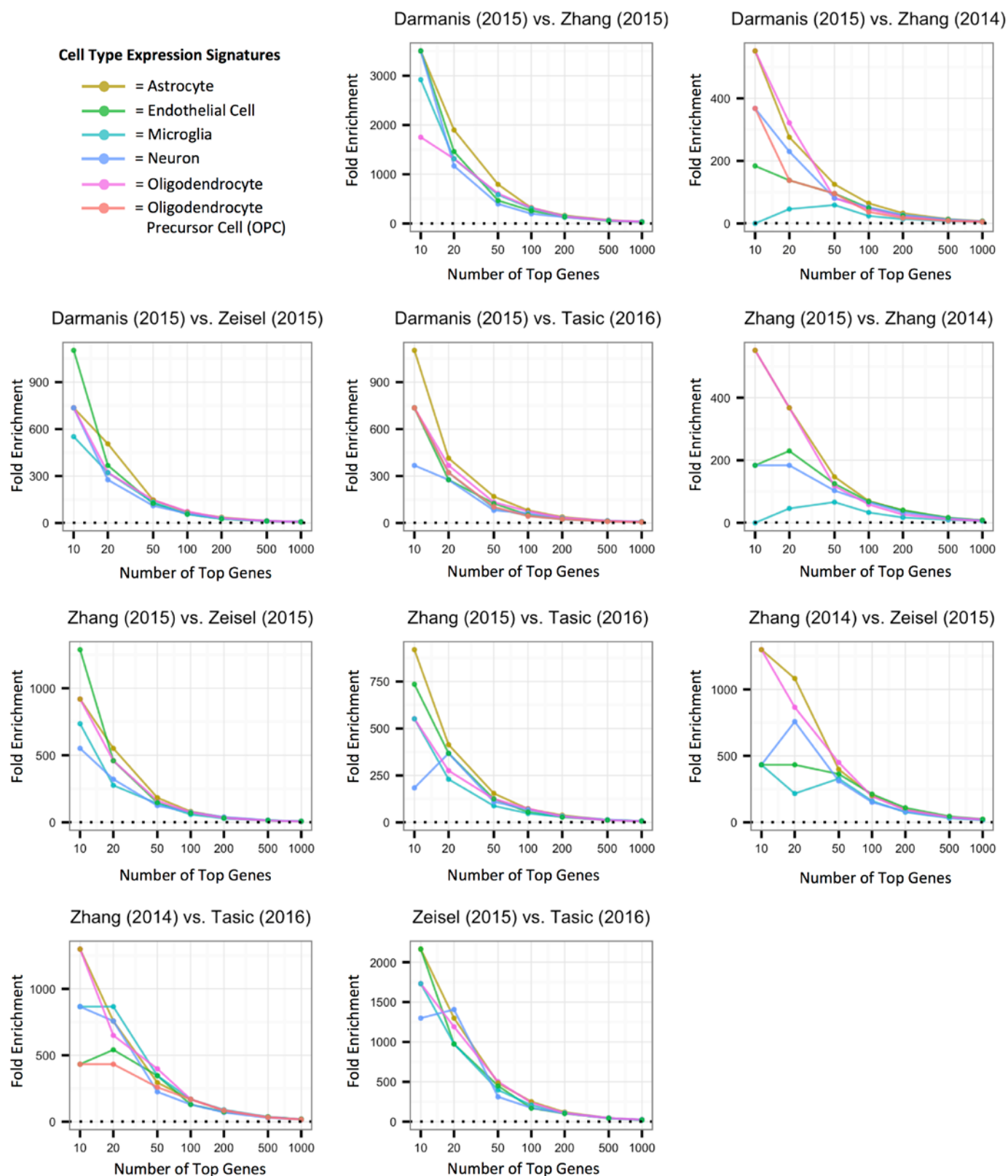
Supplemental File 4. MEGENA modules and networks identified in the human and mouse single cell RNA-seq data sets.

The Excel file contains tab-separated tables with networks and modules identified by MEGENA within each major brain cell type (astrocyte, endothelial cell, microglia, neuron, oligodendrocyte, and OPC) available in each of the Darmanis *et al.*, Tasic *et al.*, and Zeisel *et al.* data sets.

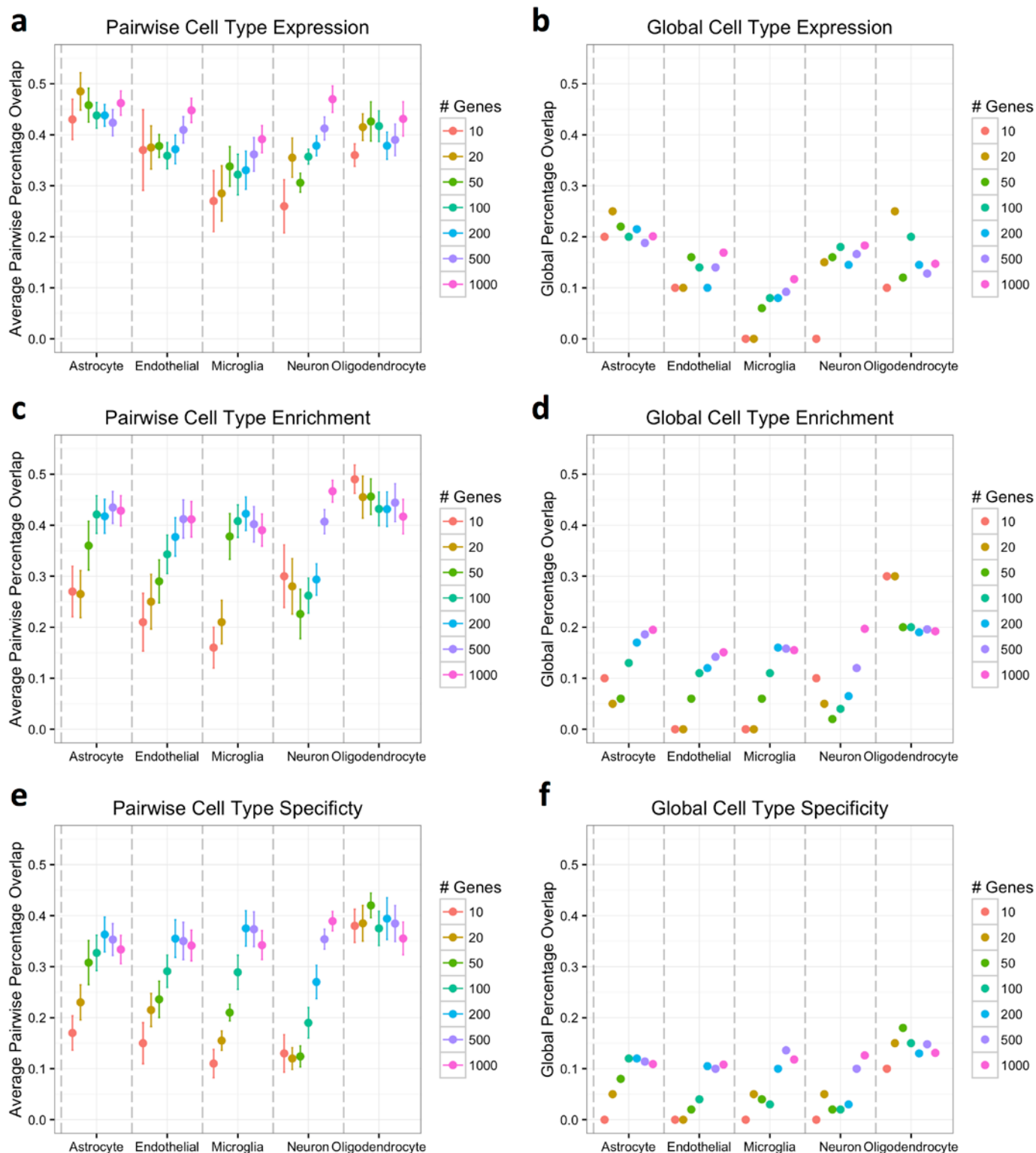
Supplementary Figures



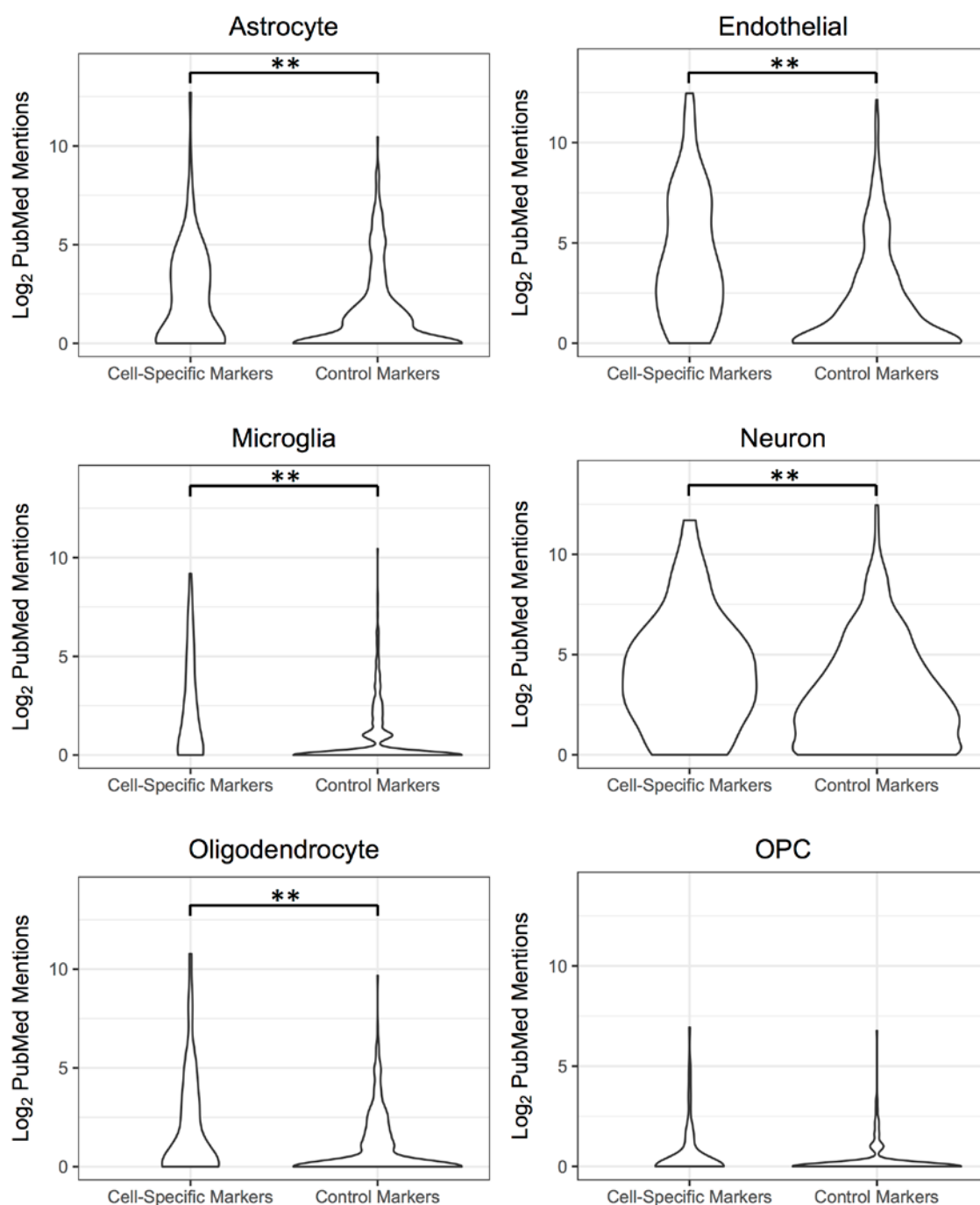
Supplementary Figure 1. Pairwise data set comparison of cell type specificity rankings for each cell type. Plots of the fold-enrichment of the intersections of genes ranked in the top n genes (where $n = 10, 20, 50, 100, 200, 500, 1000$) between pairs of data sets for the cell type specificity measure. The data sets were merged to only include gene symbols common to both prior to calculating the fold enrichment score. A fold enrichment of 0 indicates that no genes were found in the intersection of those two sets of top genes.



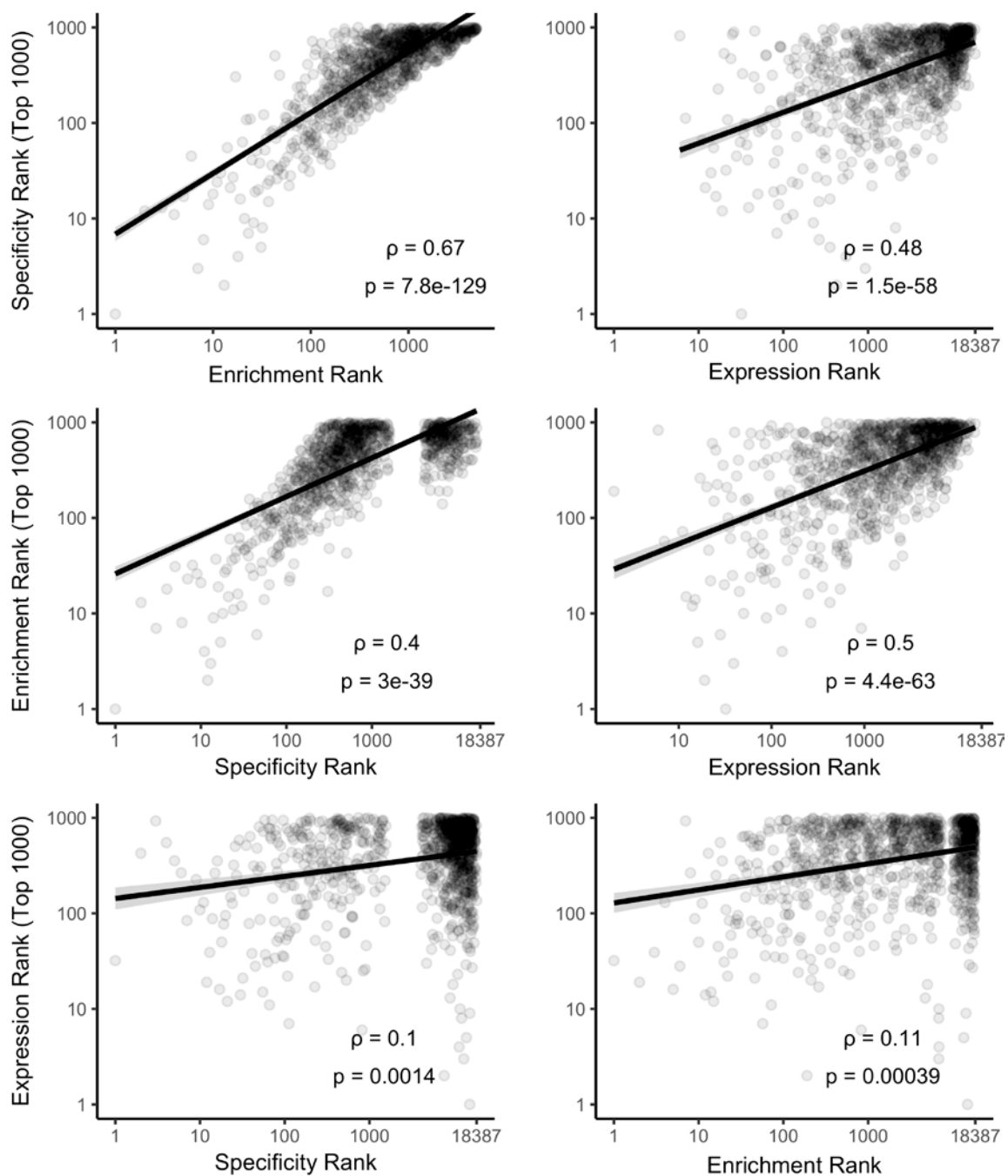
Supplementary Figure 2. Pairwise data set comparison of cell type absolute expression rankings for each cell type. Plots of the fold-enrichment of the intersections of genes ranked in the top n genes (where $n = 10, 20, 50, 100, 200, 500, 1000$) between pairs of data sets for the cell type absolute expression measure. The data sets were merged to only include gene symbols common to both prior to calculating the fold enrichment score. A fold enrichment of 0 indicates that no genes were found in the intersection of those two sets of top genes.



Supplementary Figure 3. Average percentage pairwise and global data set overlaps of top cell type-associated genes. **A, C, E:** The average percentage overlap (i.e., number of overlaps/number of top genes in each signature) of gene symbols between data sets for the top n genes in each cell type (at various values of n , i.e. $n = 10, 20, 50, 100, 200, 500, 1000$) ranked by each of the three cell type associated measures, i.e. cell type absolute expression (**A**), cell type enrichment (**C**), and cell type specificity (**E**). Dots = mean, bars = standard error of the mean. **B, D, F:** The percentage overlap of gene symbols between all five data sets for the top n genes in each cell type ranked by each of the three cell type associated measures, i.e. cell type absolute expression (**B**), cell type enrichment (**D**), and cell type specificity (**F**). Summary measures for oligodendrocyte precursor cells (OPCs) are not shown because this cell type is only present in 3 of the data sets, whereas the other cell types are represented in all 5 of the data sets.

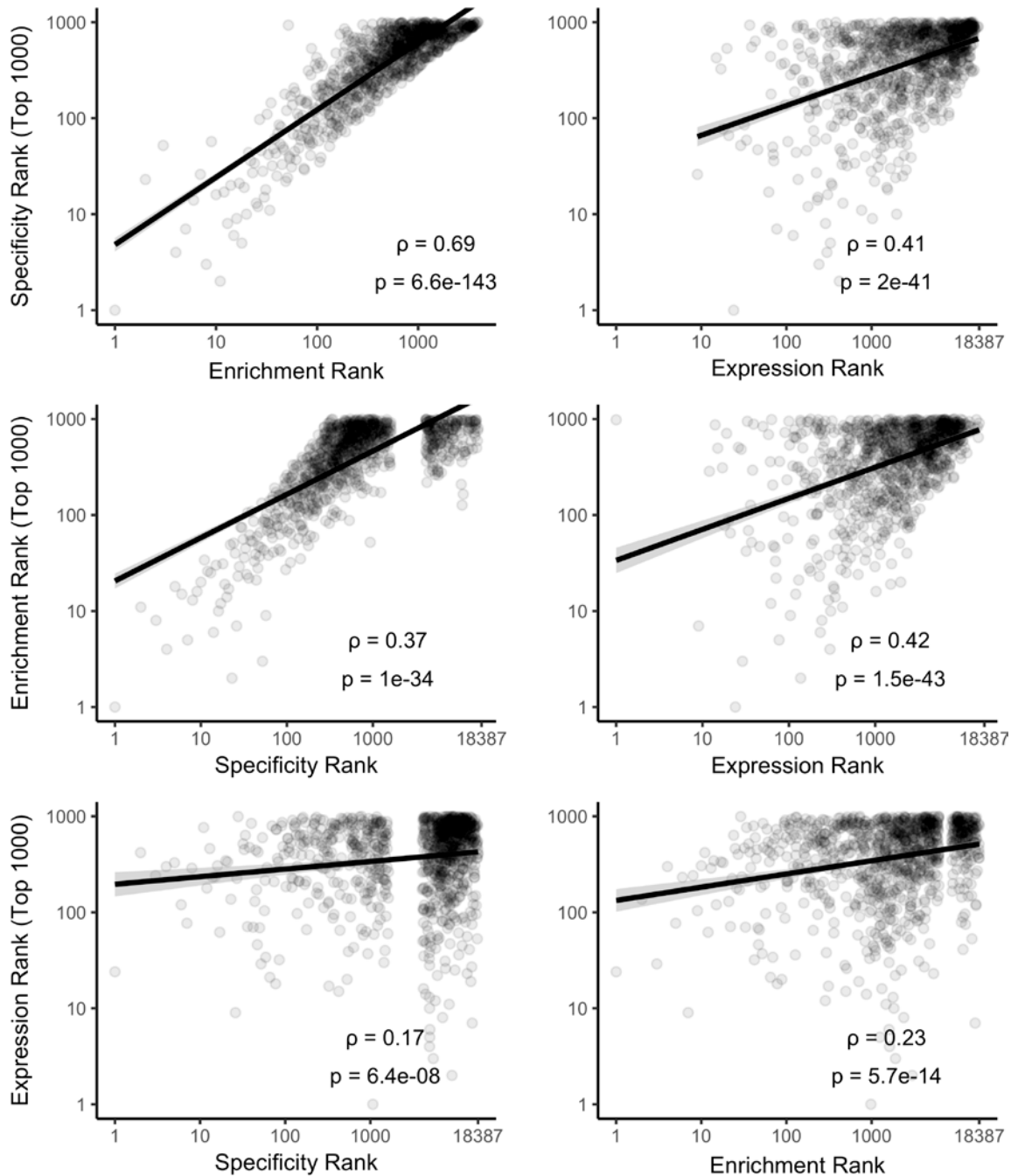


Supplementary Figure 4. PubMed text mining results of marker genes for a given cell type compared to other marker genes. For each of the six cell types, i.e. astrocyte (**A**), endothelial cell (**B**), microglia (**C**), neuron (**D**), oligodendrocyte (**E**), and oligodendrocyte precursor cell (**F**), a violin plot showing the number of PubMed abstracts that mention both that gene symbol as well as the corresponding cell type for the top 100 most enriched gene symbols in that cell type as well as the 500 other marker genes corresponding to each of the other cell types. A Mann-Whitney U test was used to compare the number of PubMed mentions for cell-specific markers compared to the number of PubMed mentions for the control markers ((*, $p < 0.05$; **, $p < 0.01$).

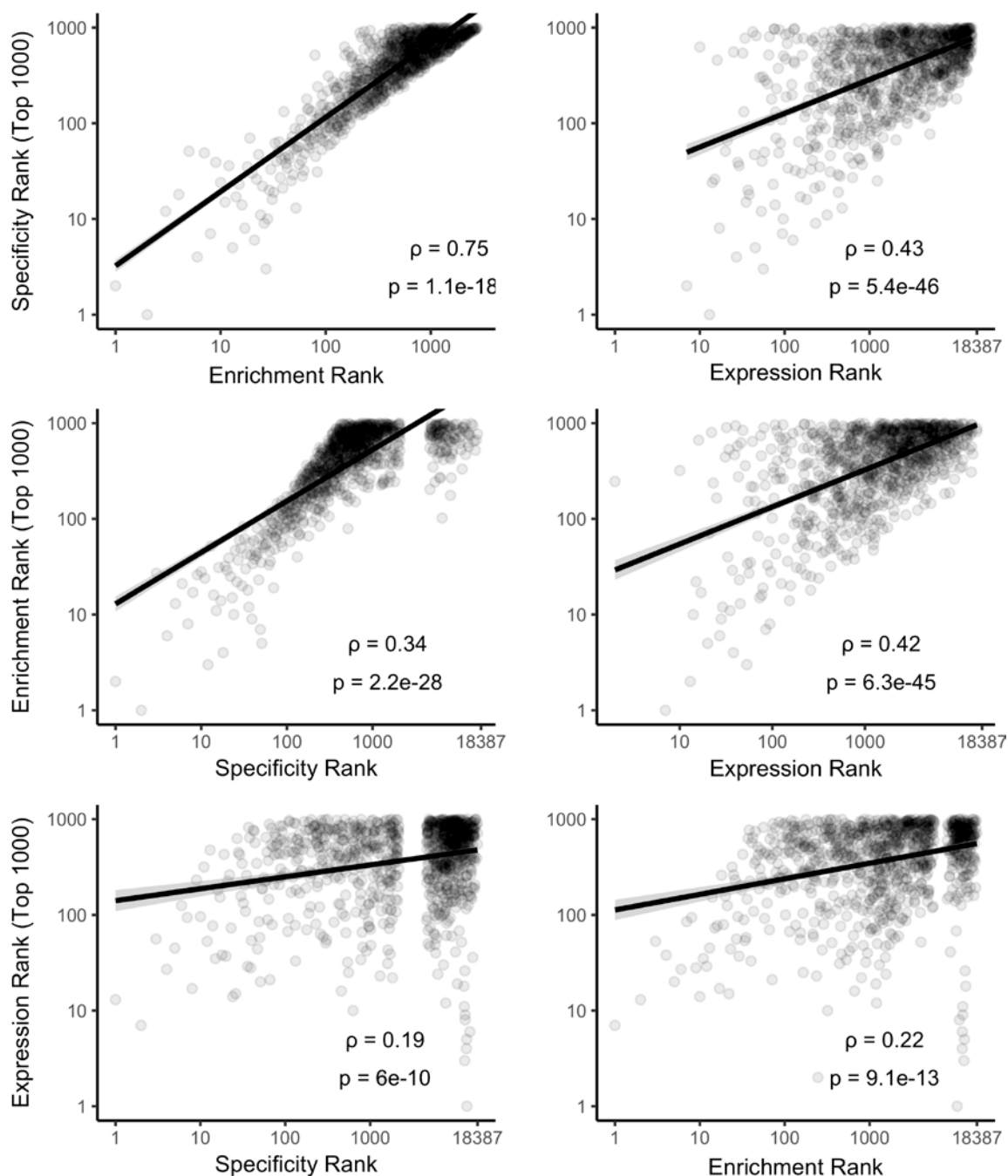


Supplementary Figure 5. Comparison of the three cell type-associated measures in astrocytes. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the ranks of one gene according to each measure. The points are partially transparent ($\alpha = 0.1$) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95%

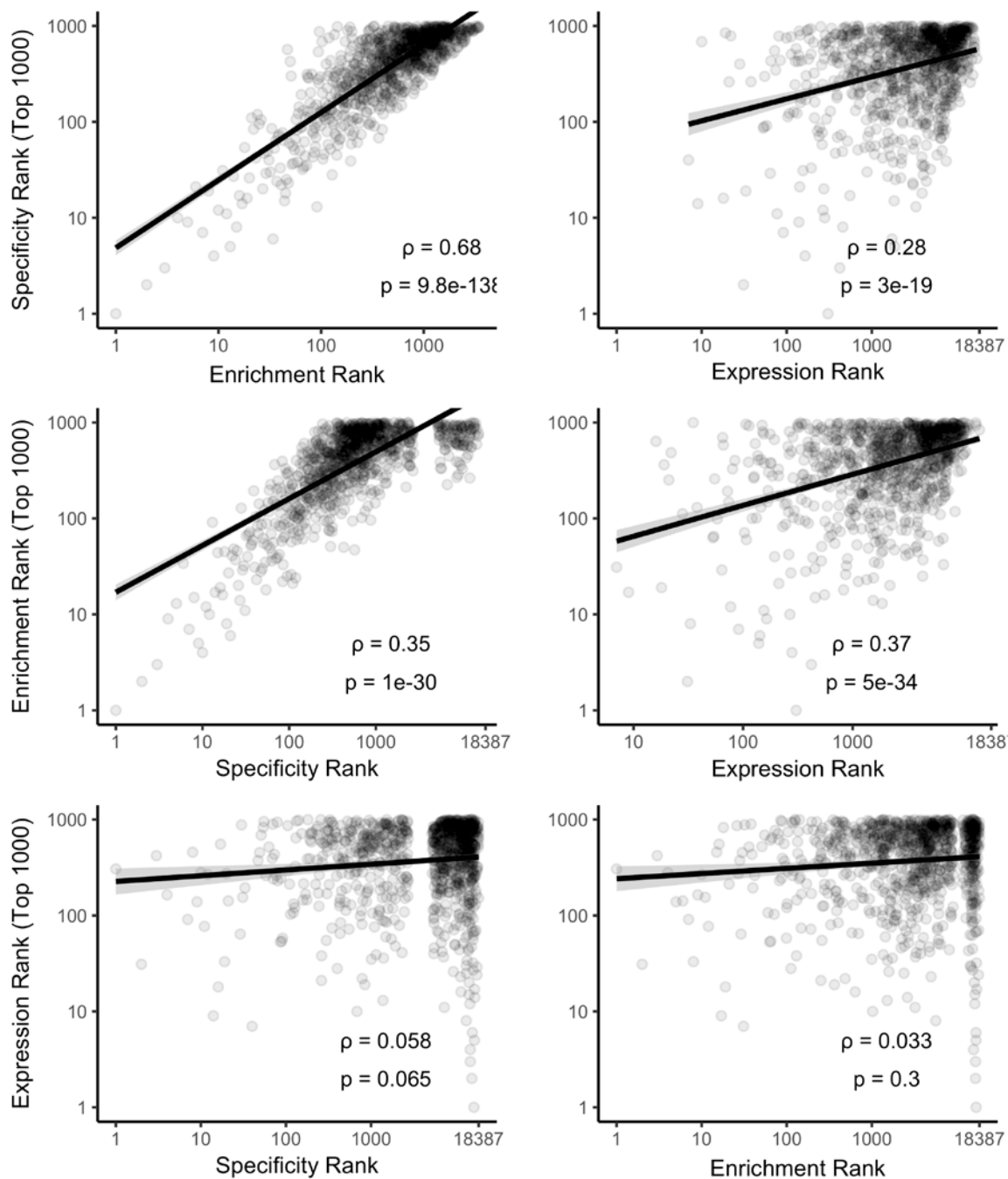
confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-value for each comparison are noted in the bottom right of each plot.



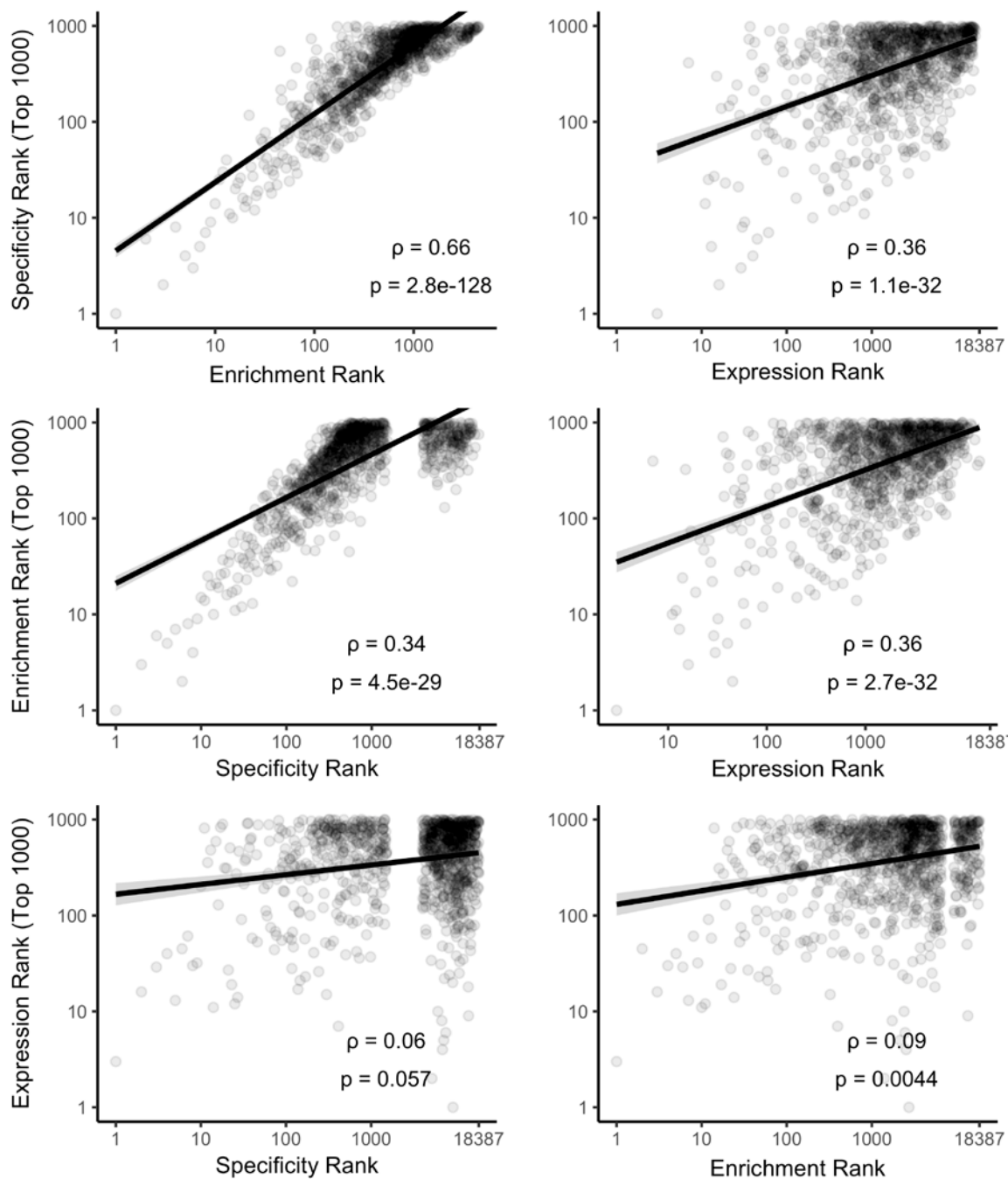
Supplementary Figure 6. Comparison of the three cell type-associated measures in endothelial cells. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the rank of one gene according to each measure. The points are partially transparent ($\alpha = 0.1$) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-values for each comparison are noted in the bottom right of each plot.



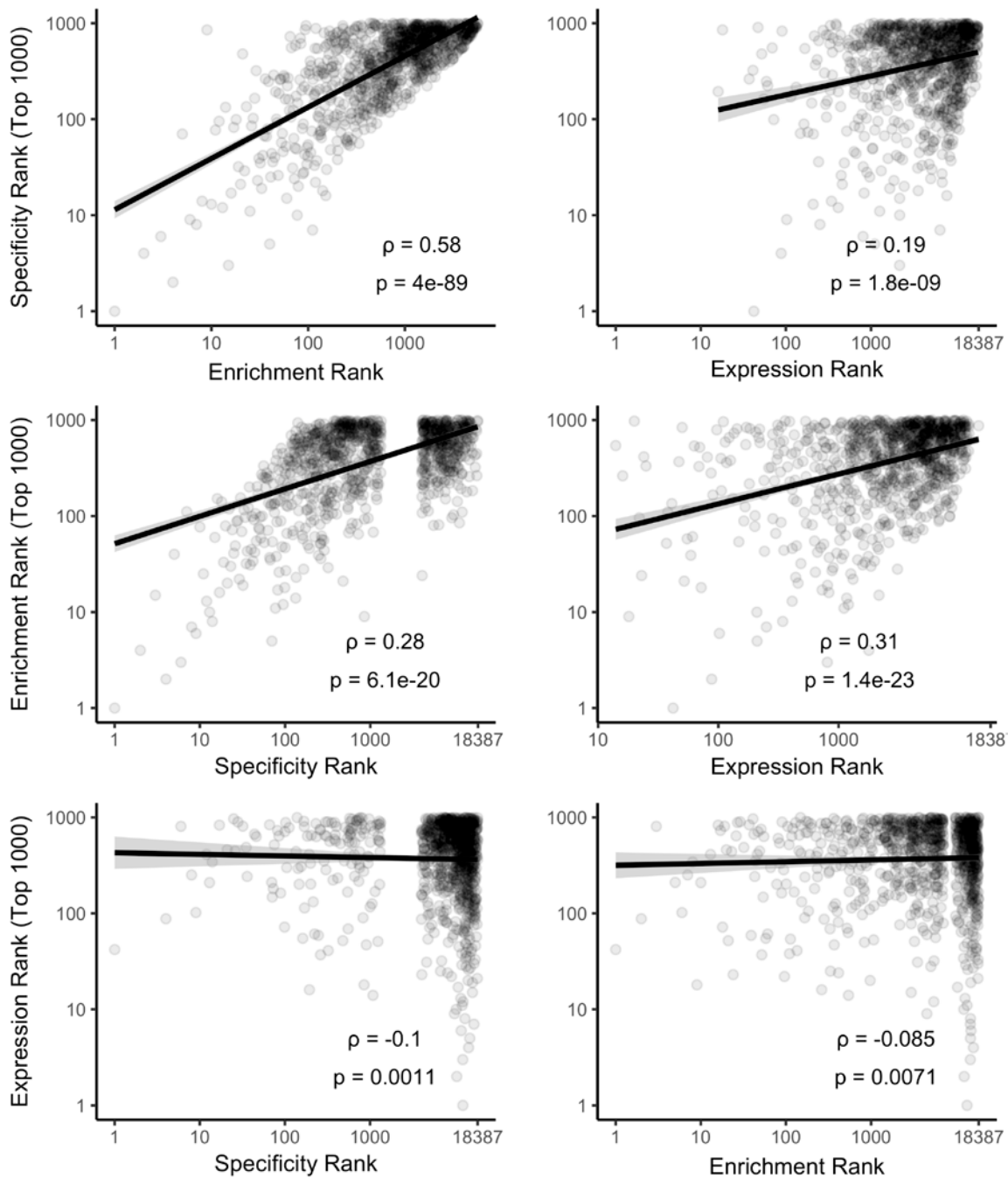
Supplementary Figure 7. Comparison of the three cell type-associated measures in microglia. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the rank of one gene according to each measure. The points are partially transparent ($\alpha = 0.1$) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-values for each comparison are noted in the bottom right of each plot.



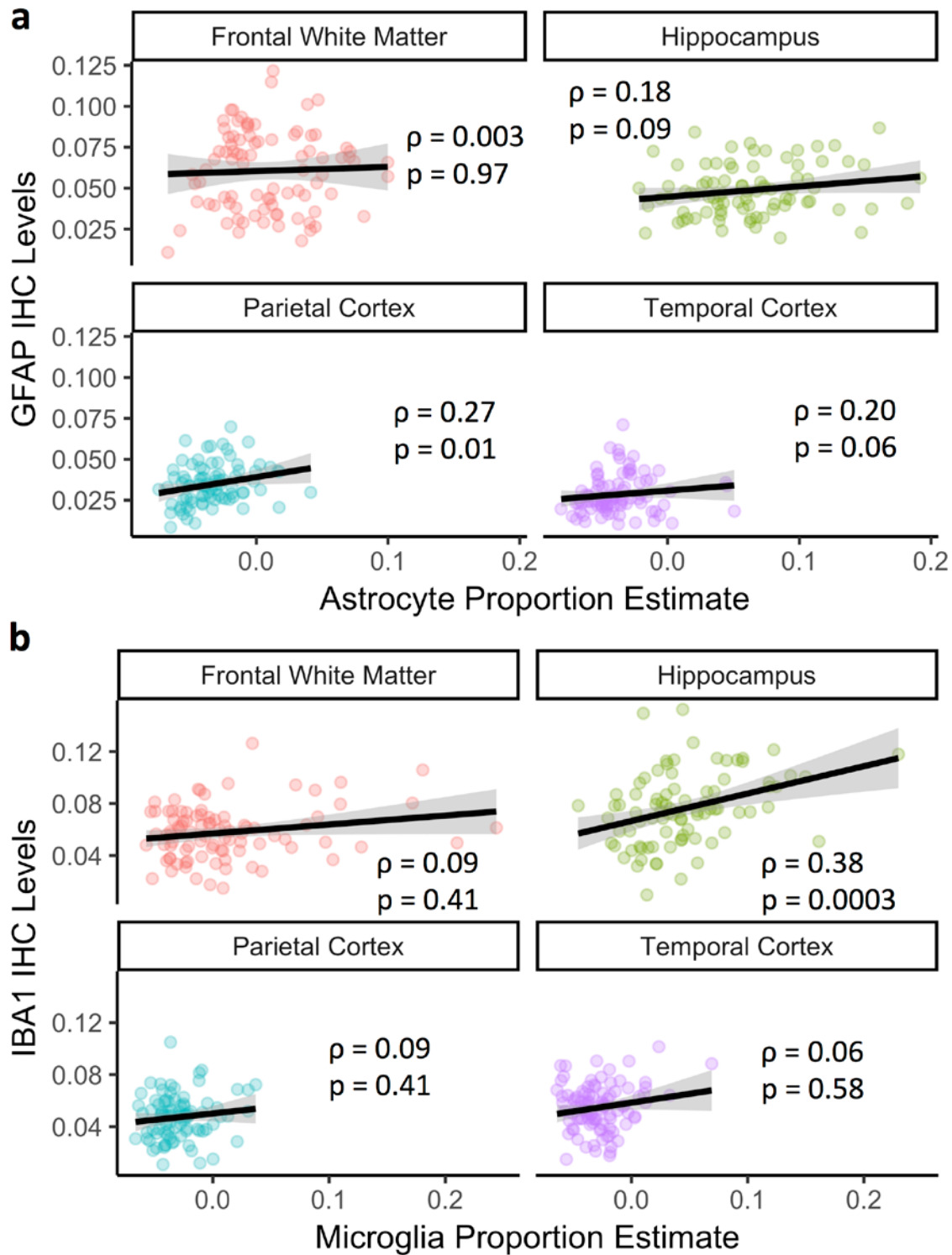
Supplementary Figure 8. Comparison of the three cell type-associated measures in neurons. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the rank of one gene according to each measure. The points are partially transparent (alpha = 0.1) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-values for each comparison are noted in the bottom right of each plot.



Supplementary Figure 9. Comparison of the three cell type-associated measures in oligodendrocytes. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the rank of one gene according to each measure. The points are partially transparent ($\alpha = 0.1$) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-values for each comparison are noted in the bottom right of each plot.

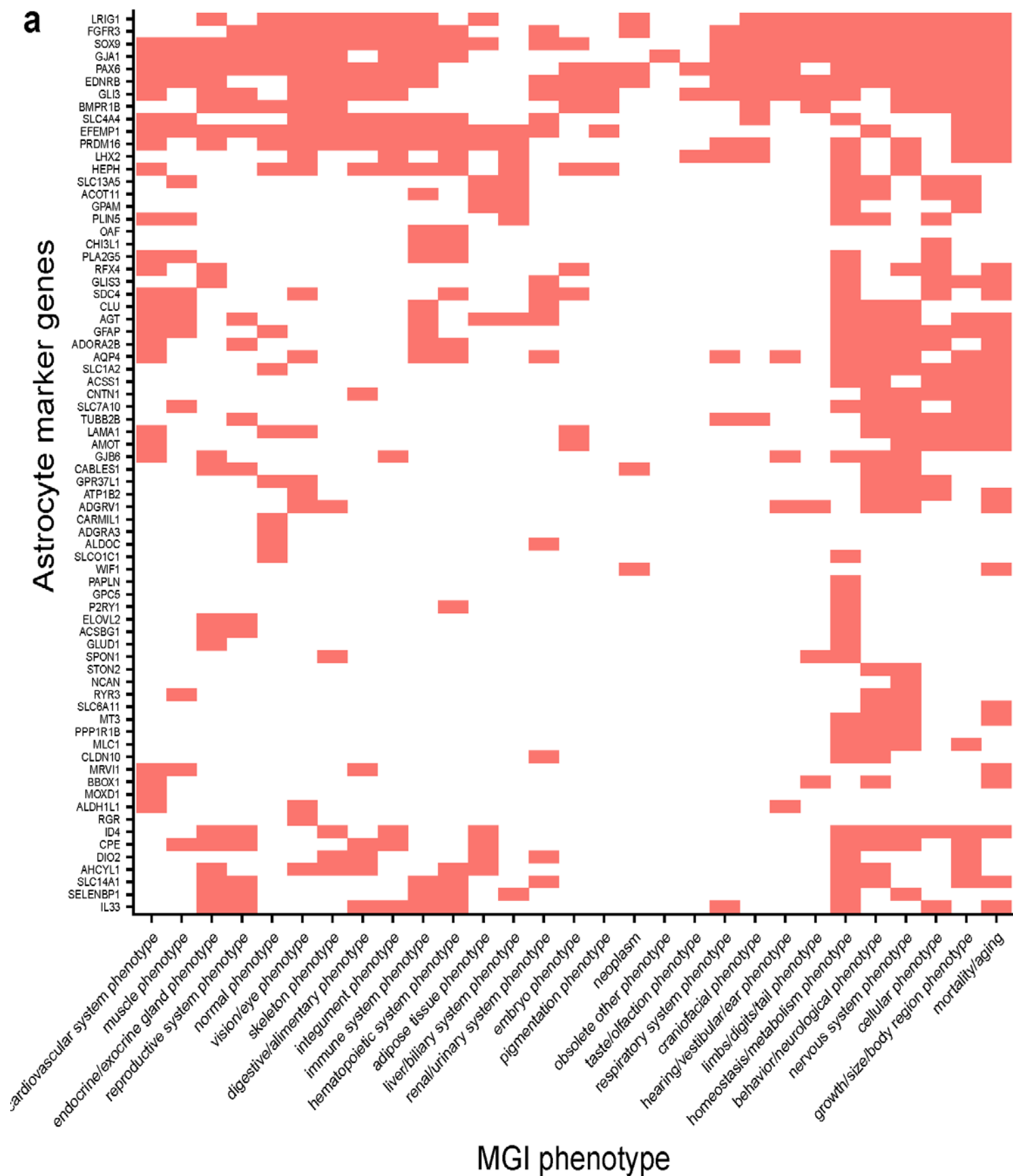


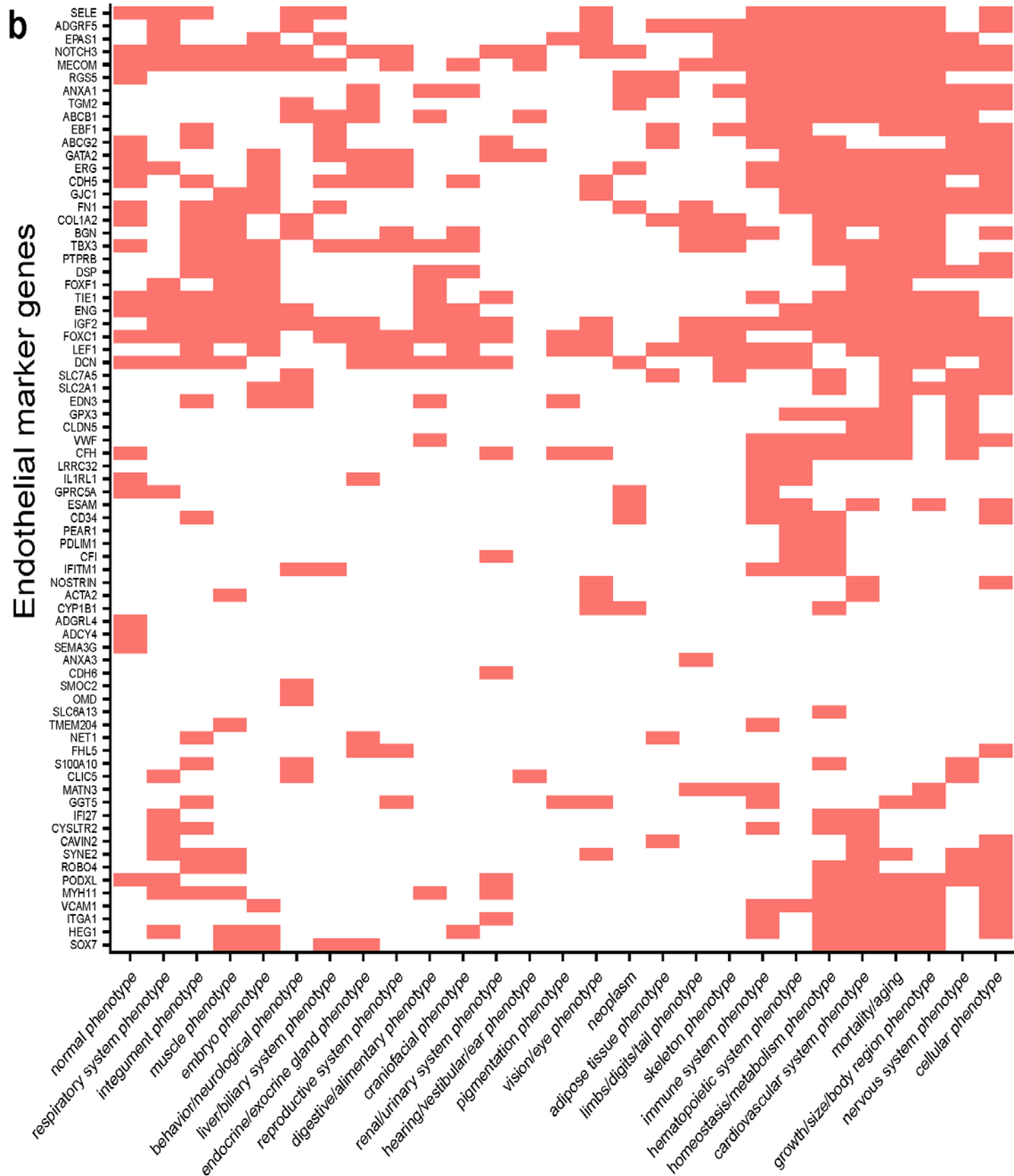
Supplementary Figure 10. Comparison of the three cell type-associated measures in oligodendrocyte precursor cells. The ranks of the top 1000 genes in each of the three cell type-associated measures (y-axis) compared to the ranks of the other two measures (x-axes; see the Supplementary Methods for more information). Each point represents the rank of one gene according to each measure. The points are partially transparent ($\alpha = 0.1$) to mitigate overplotting. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. Note that the discontinuities present in the plots of specificity and enrichment rankings are due to the set of genes with no measured expression in this cell type. The rank correlation and associated p-values for each comparison are noted in the bottom right of each plot.

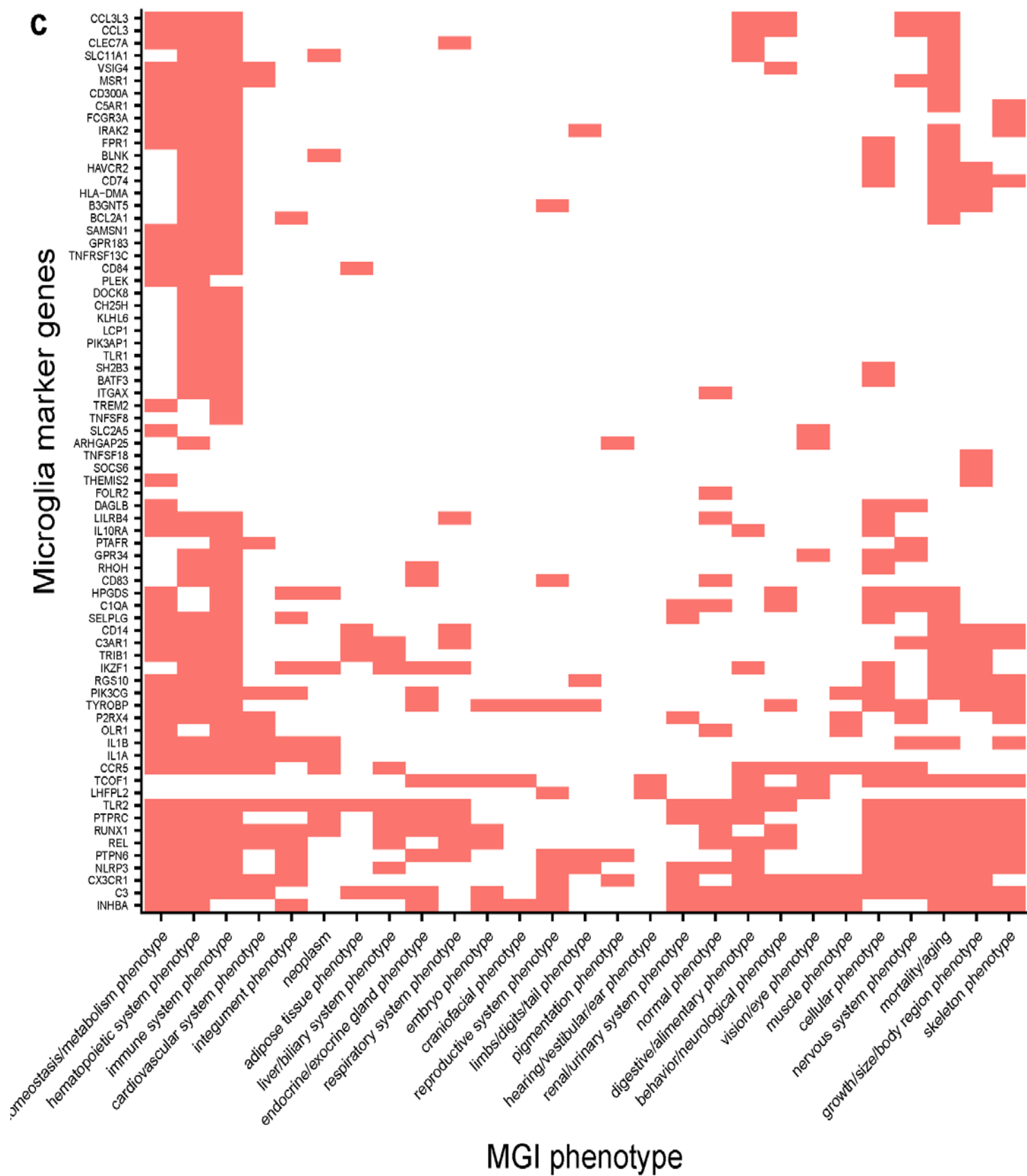


Supplementary Figure 11. Within-region rank correlation between estimated relative cell type proportion and IHC quantifications for astrocytes and microglia.

For each of the four brain regions present in the Allen Brain Atlas Aging, Dementia, and TBI study, we estimated the relative proportion of astrocytes and microglia using an SVD-based method on the top 50 marker genes (x-axis) and found the rank correlation of these estimates with the immunohistochemistry (IHC) quantifications for GFAP and IBA1 (y-axis), respectively. The black line is a result of a linear model fit to the data, while the grey lines represent 95% confidence intervals. FWM, frontal white matter; HIP, hippocampus; PCx, parietal cortex; TCx, temporal cortex.

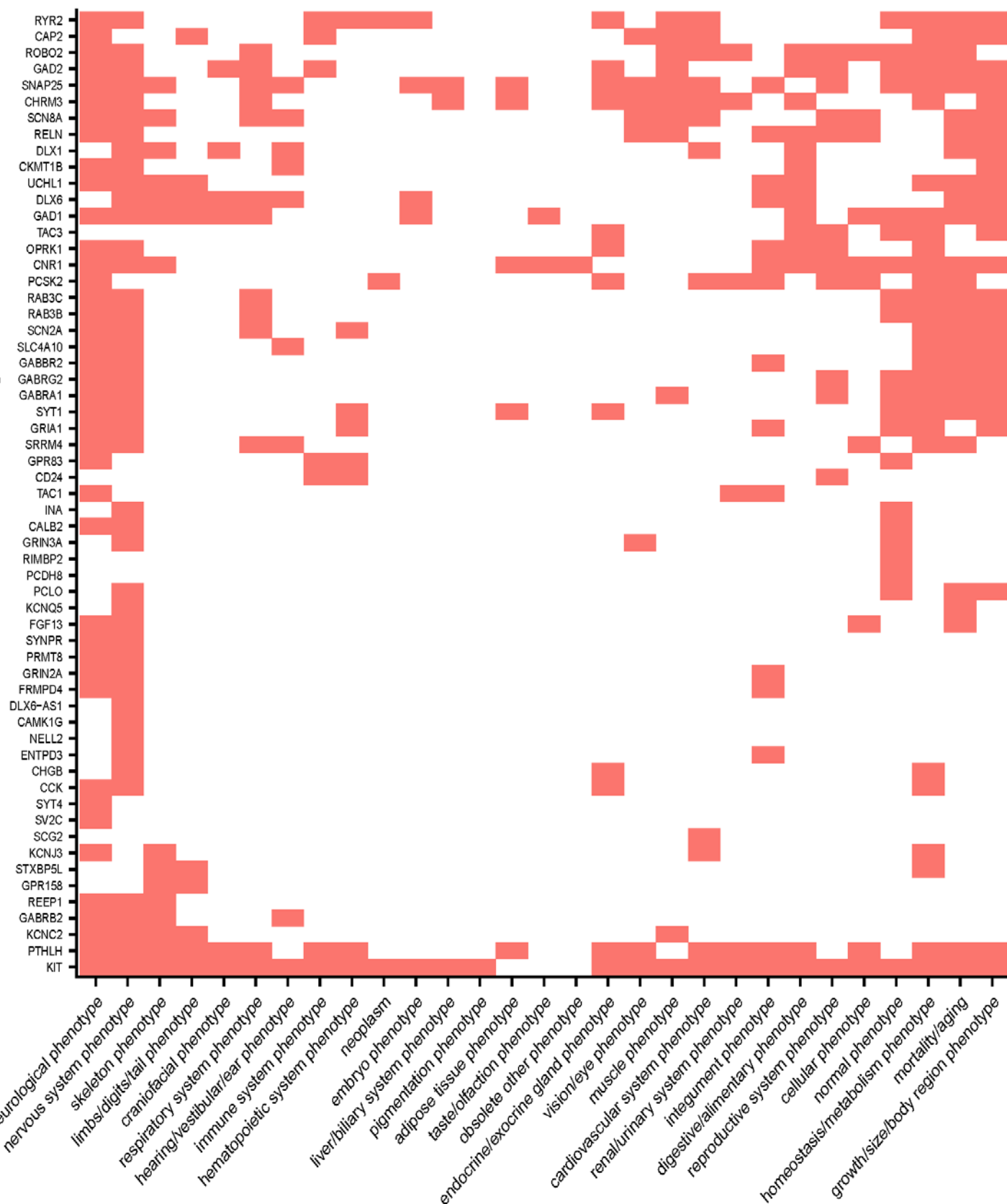




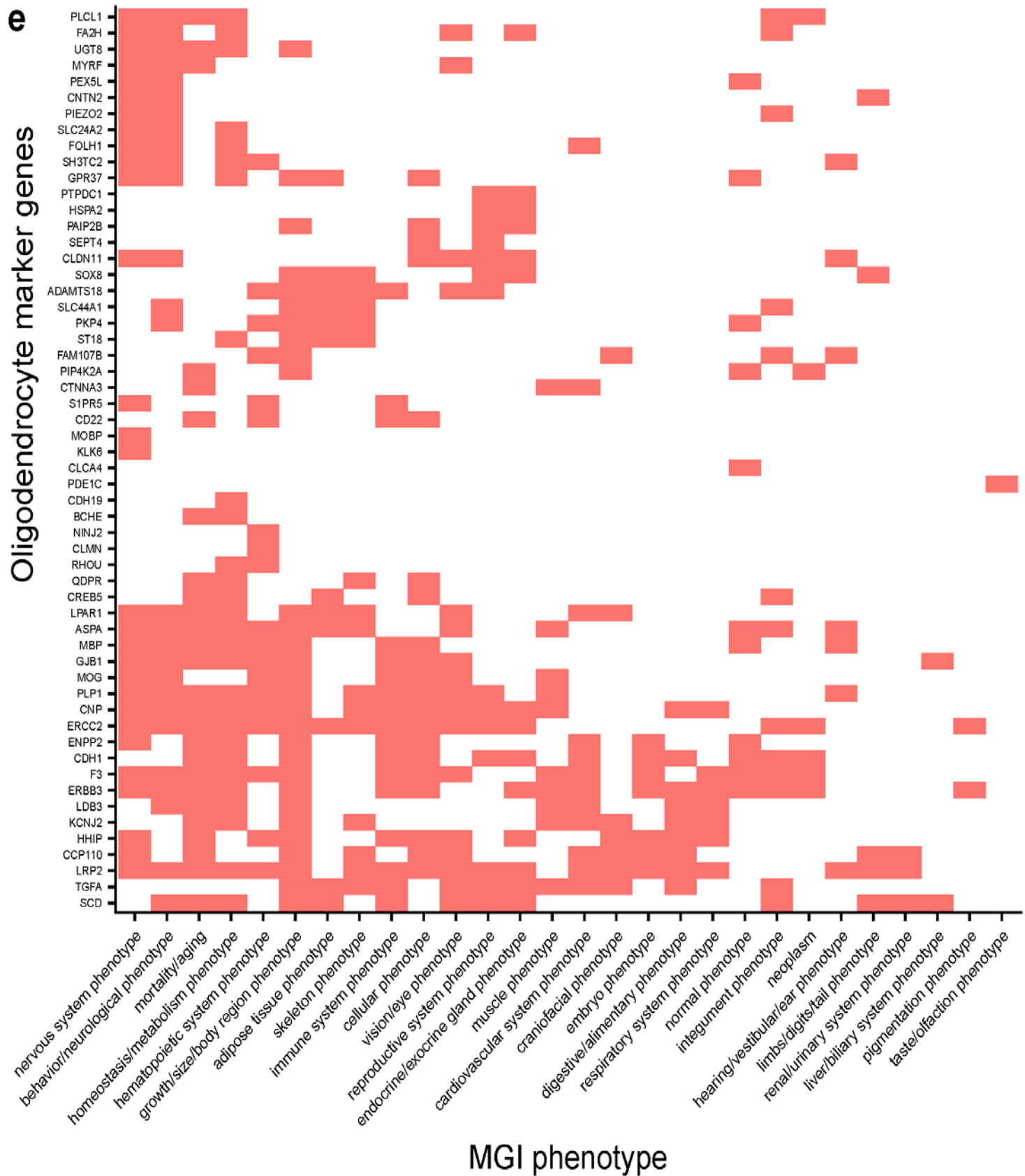


d

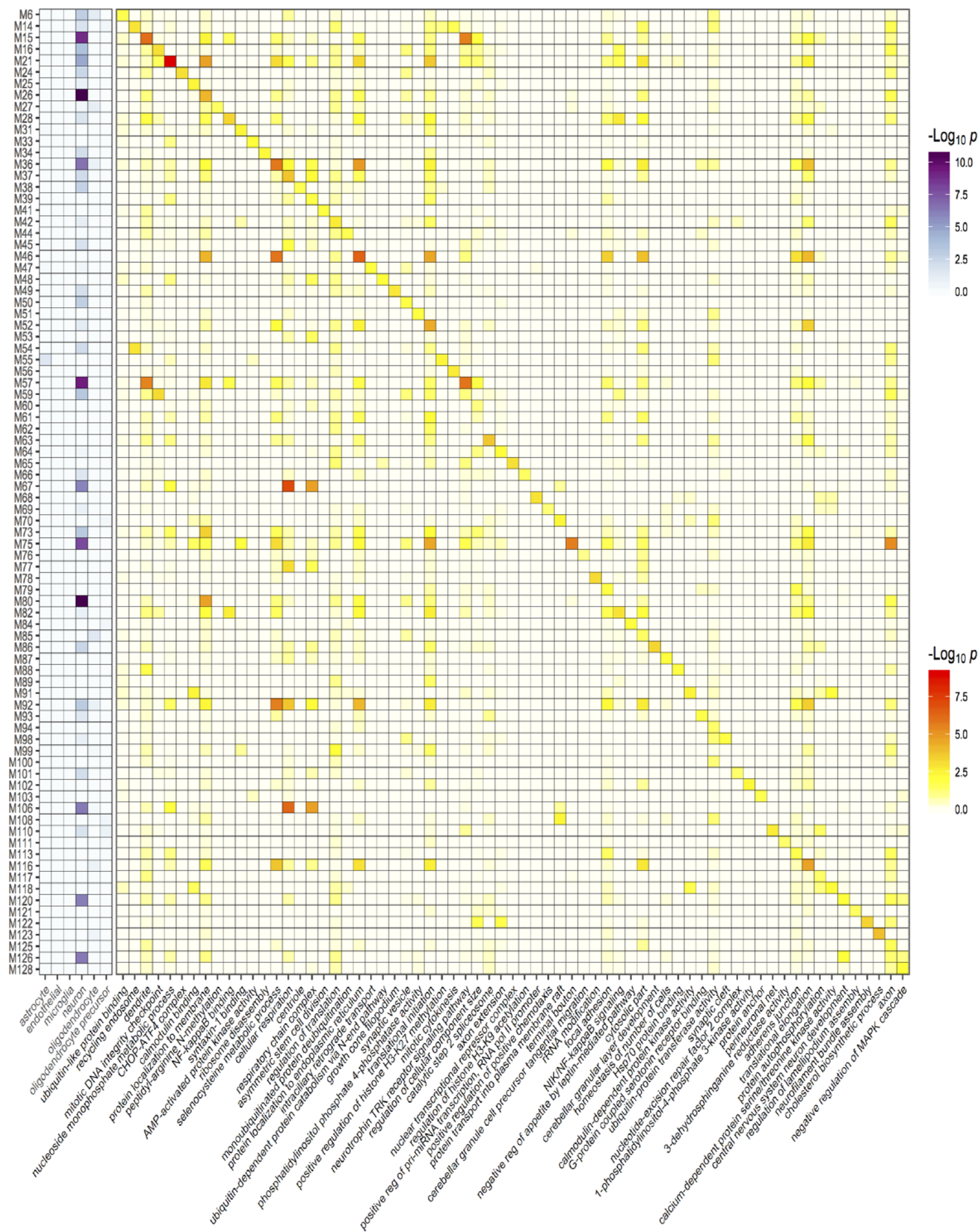
Neuron marker genes



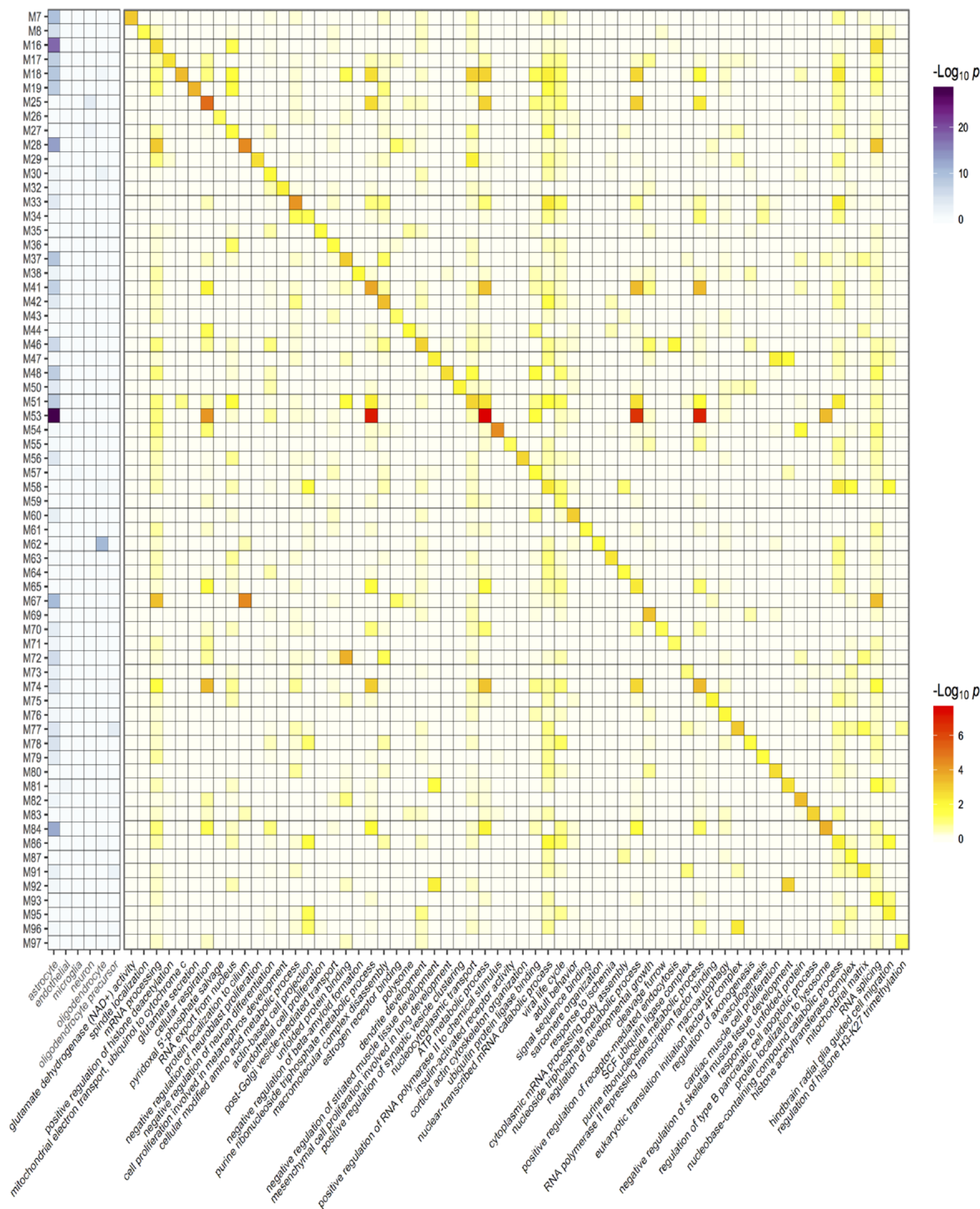
MGI phenotype



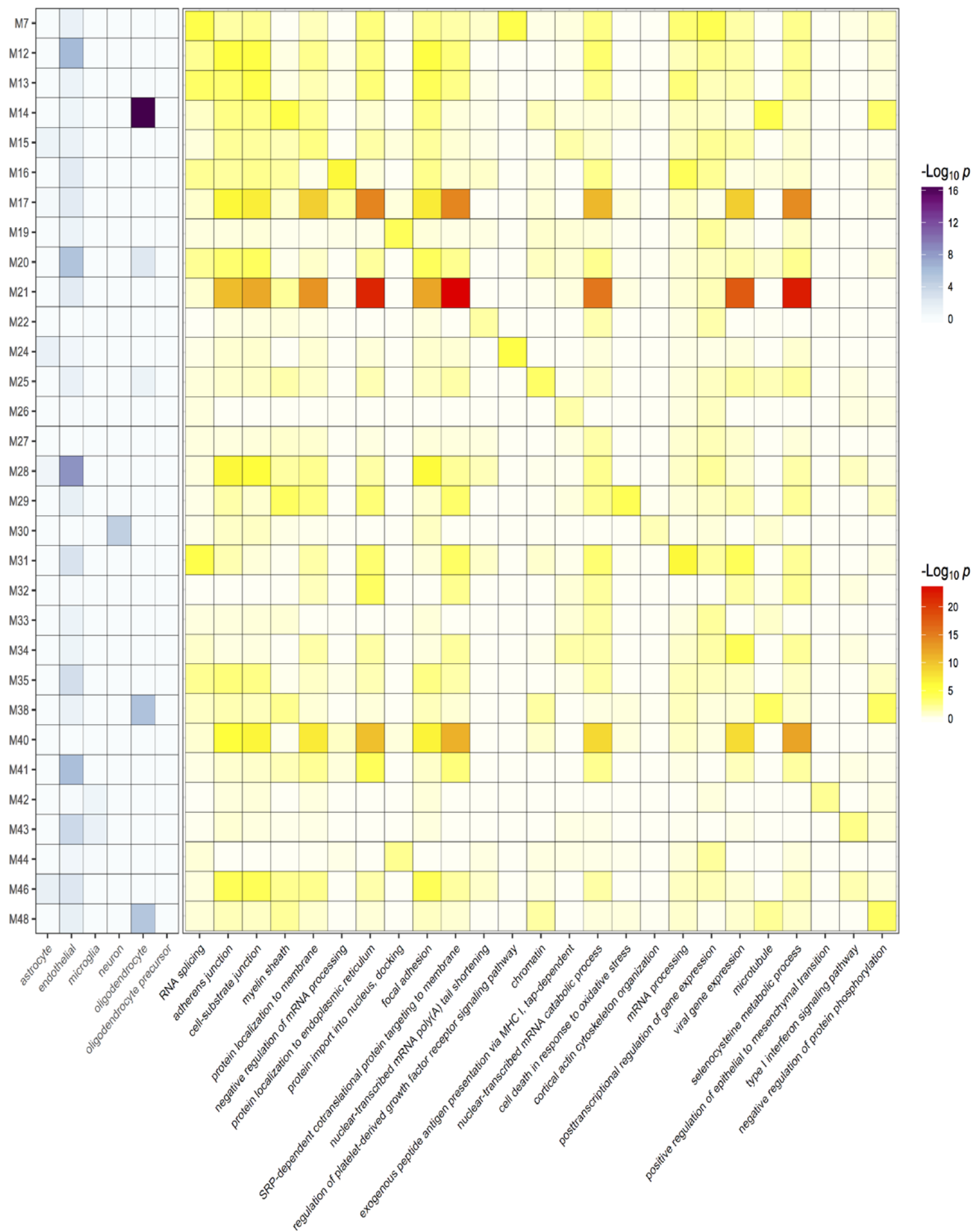
Supplementary Figure 12. MGI phenotypes associated with the brain cell type marker genes.



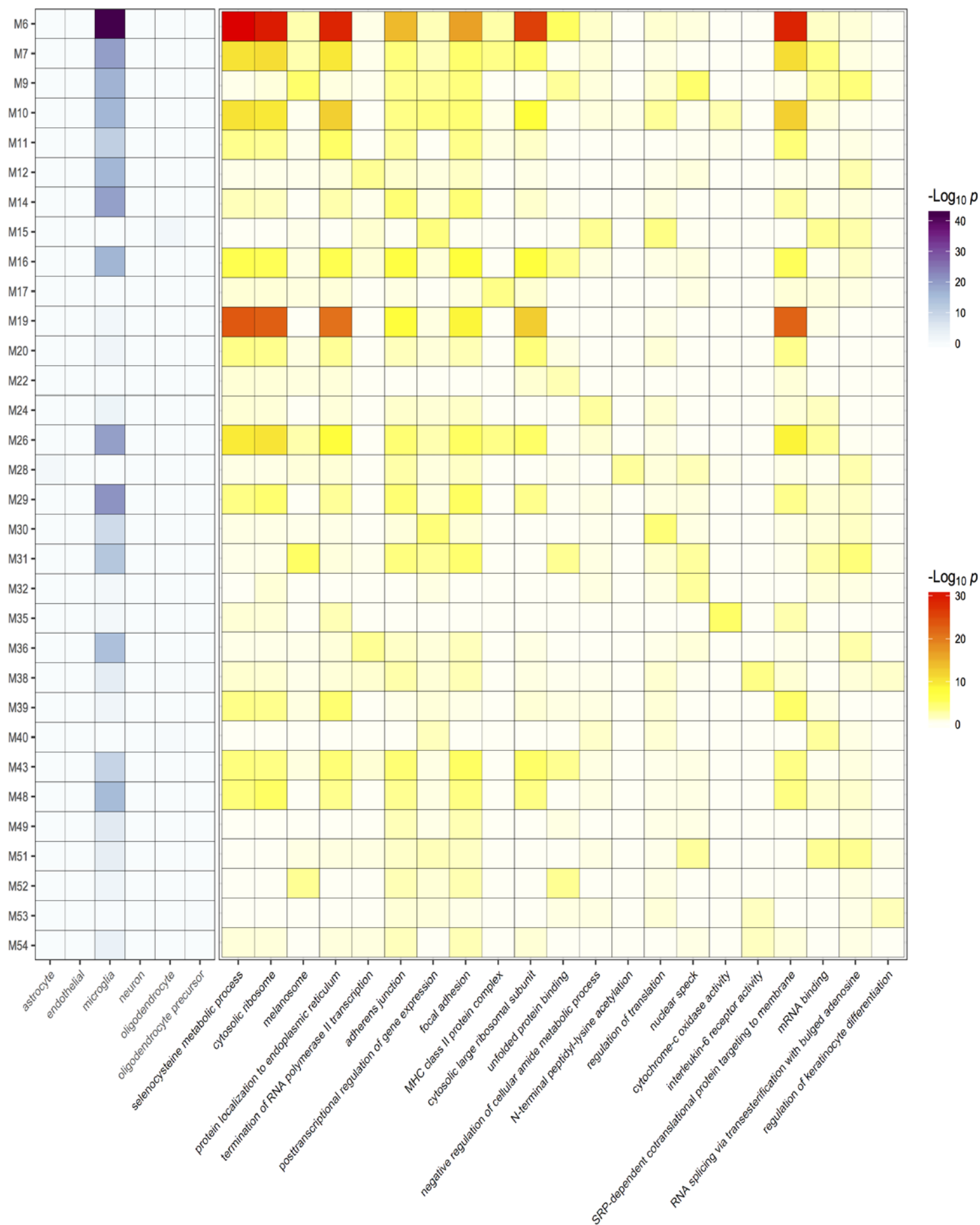
Supplementary Figure 13. Enrichments of MEGENA modules from human neurons in brain cell type signatures and GO terms. The modules on the left match the modules in the multiscale network shown in **Figure 8A**. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.



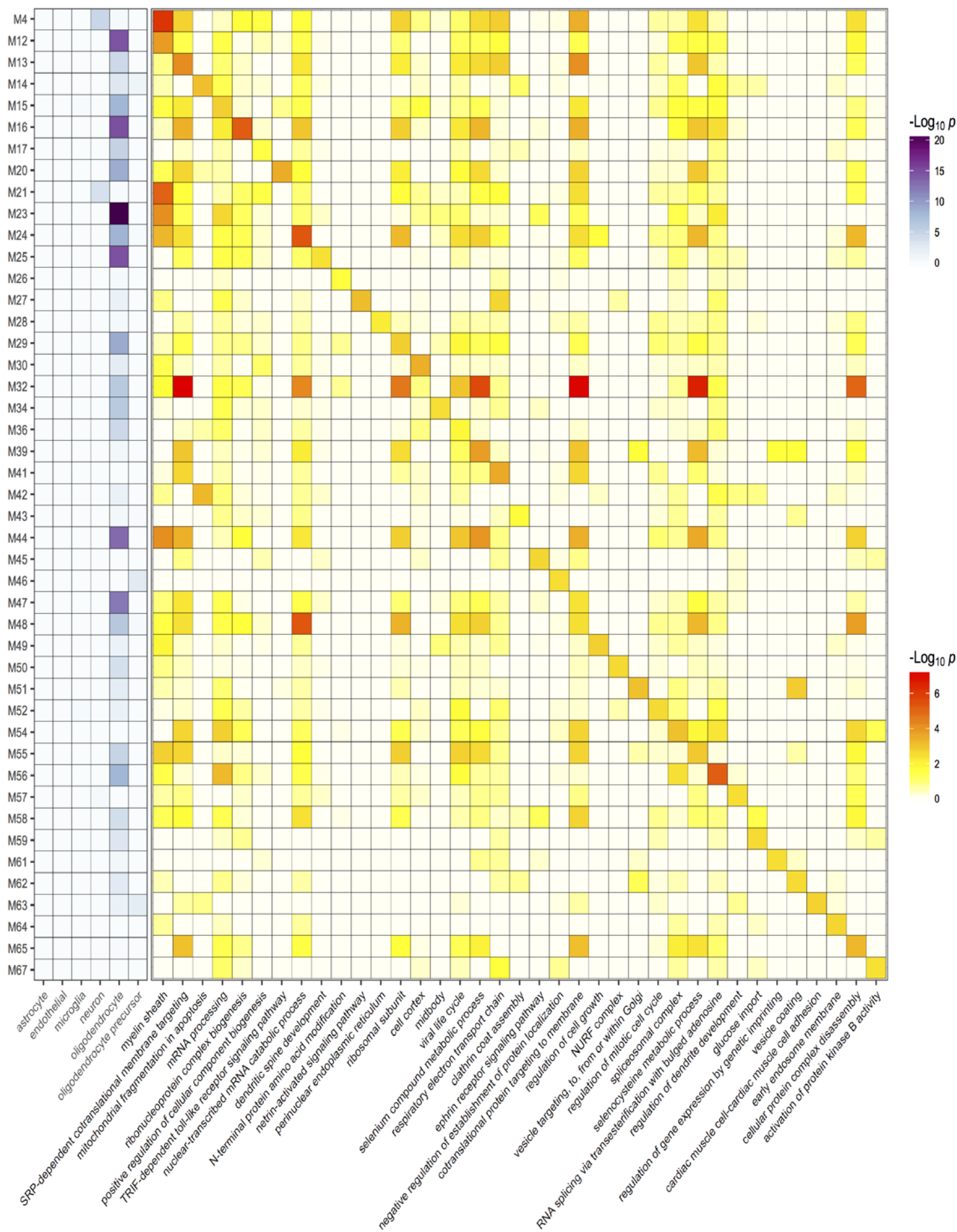
Supplementary Figure 14. Enrichments of MEGENA modules from human astrocytes in brain cell type signatures and GO terms. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.



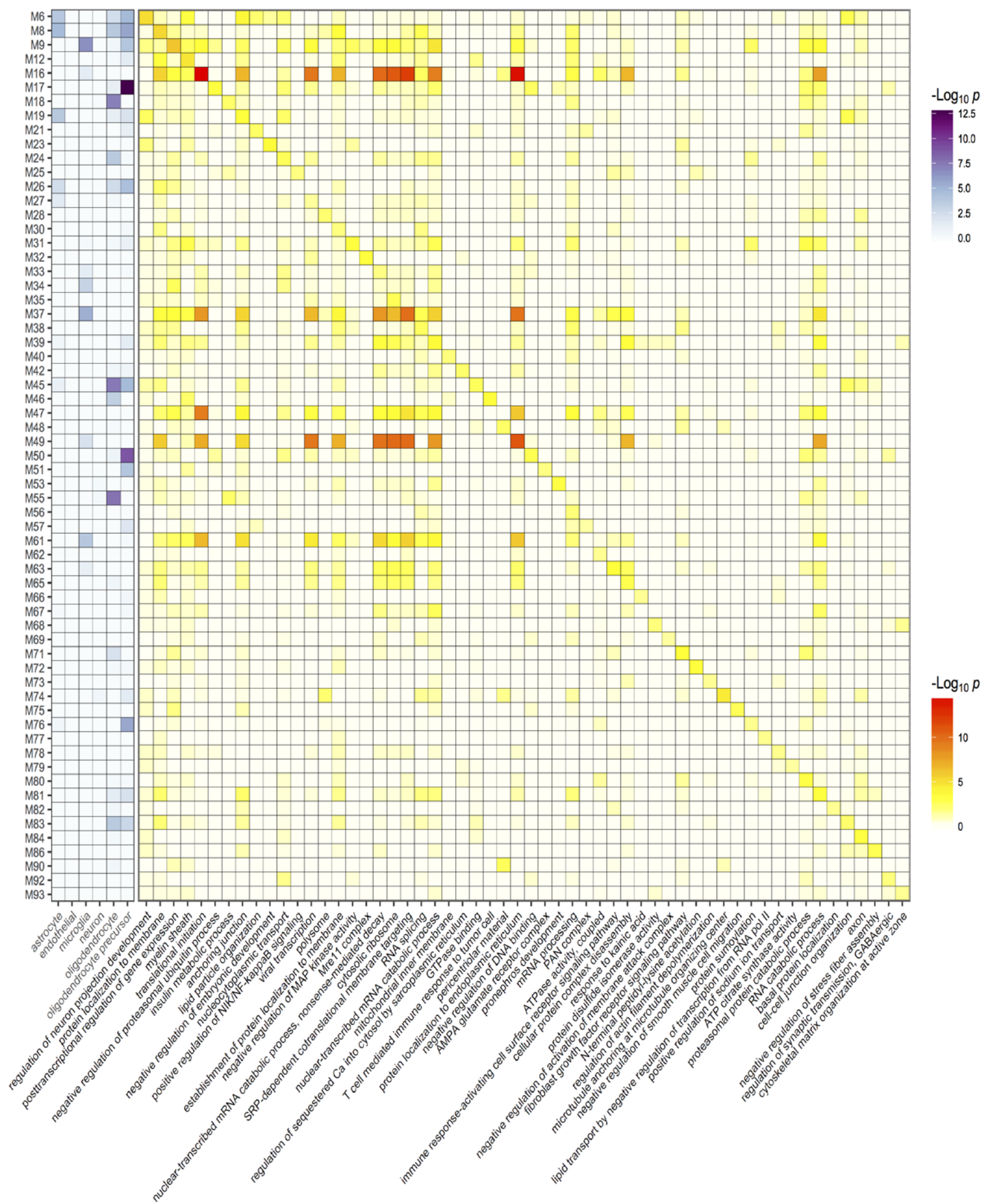
Supplementary Figure 15. Enrichments of MEGENA modules from human endothelial cells in brain cell type signatures and GO terms. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.



Supplementary Figure 16. Enrichments of MEGENA modules from human microglia in brain cell type signatures and GO terms. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.



Supplementary Figure 17. Enrichments of MEGENA modules from human oligodendrocytes in brain cell type signatures and GO terms. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.



Supplementary Figure 18. Enrichments of MEGENA modules from human oligodendrocyte precursor cells in brain cell type signatures and GO terms. The left panel shows the enrichment (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) of each module in the top 500 most cell type-specific genes for each cell type in humans that we identified in this manuscript. The right panel shows the most significantly enriched GO term (Benjamini-Hochberg adjusted $-\log_{10}$ p-value) for the genes in each of the corresponding modules, along with the GO enrichment of that same GO term in all the other modules.

References

- 1 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2012).
- 2 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).
- 3 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
- 4 McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *bioRxiv*, doi:10.1101/069633 (2016).
- 5 Bullard, J. H. *et al.* Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94, doi:10.1186/1471-2105-11-94 (2010).
- 6 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)* **21**, 3439-3440, doi:10.1093/bioinformatics/bti525 (2005).
- 7 Miller, J. A. *et al.* Strategies for aggregating gene expression data: the collapseRows R function. *BMC bioinformatics* **12**, 322, doi:10.1186/1471-2105-12-322 (2011).
- 8 Chikina, M., Zaslavsky, E. & Sealfon, S. C. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* **31**, 1584-1591, doi:10.1093/bioinformatics/btv015 (2015).
- 9 Miller, J. A. *et al.* Neuropathological and transcriptomic characteristics of the aged brain. *Elife* **6**, doi:10.7554/eLife.31126 (2017).
- 10 Kozlenkov, A. *et al.* Substantial DNA methylation differences between two major neuronal subtypes in human brain. *Nucleic acids research* **44**, 2593-2612, doi:10.1093/nar/gkv1304 (2016).
- 11 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).
- 12 McKenzie, A. T., Katsyov, I., Song, W. M., Wang, M. & Zhang, B. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst Biol* **10**, 106, doi:10.1186/s12918-016-0349-1 (2016).
- 13 Falcon, S. & Gentleman, R. Using GOSTATS to test gene lists for GO term association. *Bioinformatics (Oxford, England)* **23**, 257-258, doi:10.1093/bioinformatics/btl567 (2007).