**Supplemental Information**

# Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets

Emre Caglayan, Yuxiang Liu and Genevieve Konopka

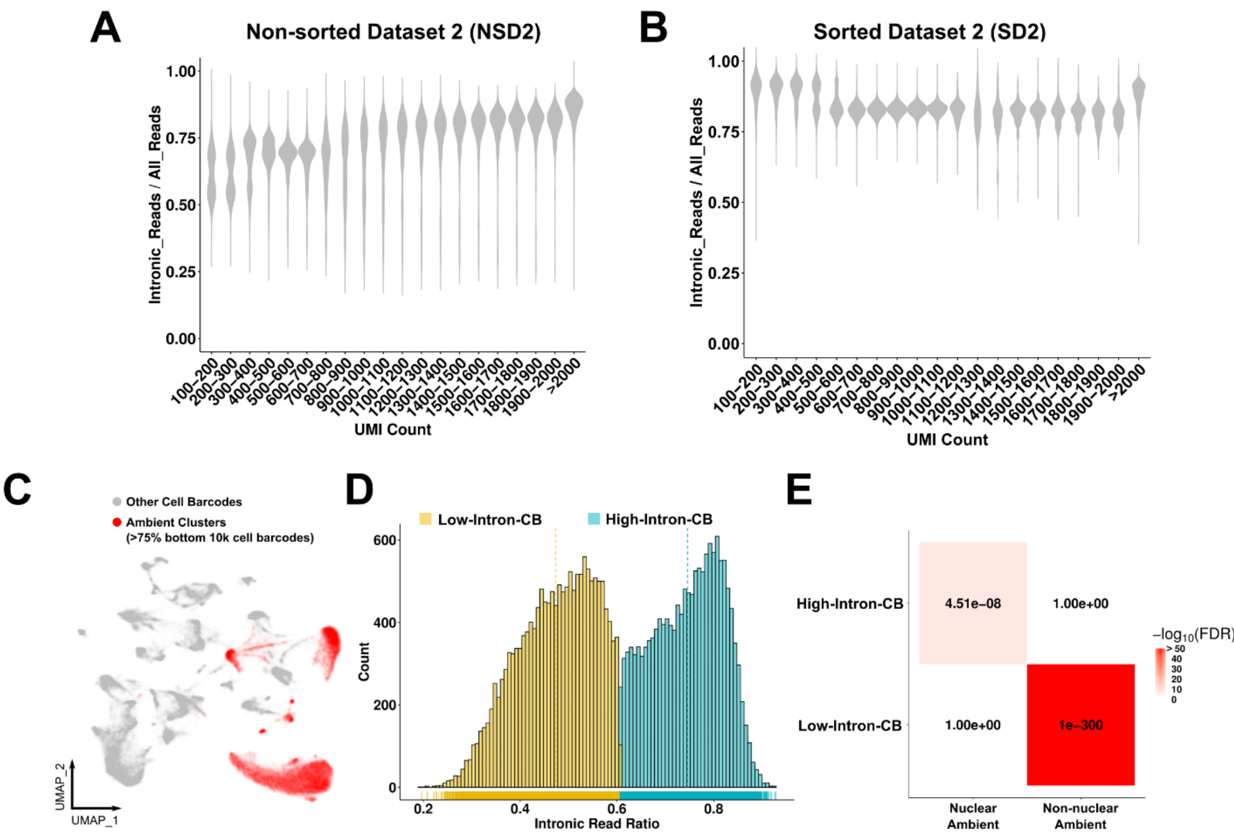


**Figure S1 (Related to Figure 2). Independent confirmation of nuclear and non-nuclear ambient RNA types. (A-B)** Plots of intronic read ratios across increasing UMI counts in **(A)** non-sorted dataset 2 (NSD2) and **(B)** sorted dataset 2 (SD2). UMI counts are divided into intervals of 100 from 100-2000. **(C)** The clusters that contain >75% of filtered cell barcodes are highlighted and named ambient clusters (dataset: NSD2). **(D)**. Plot of the distribution of intronic read ratios within ambient clusters. Yellow: Low-Intron-CB (CB: Cell Barcodes), blue: High-Intron-CB. **(E)** Heatmap of enrichments between ambient RNA types. Nuclear ambient RNA and non-nuclear ambient RNAs from SD1 and NSD1 were compared (see Figures 1-2).

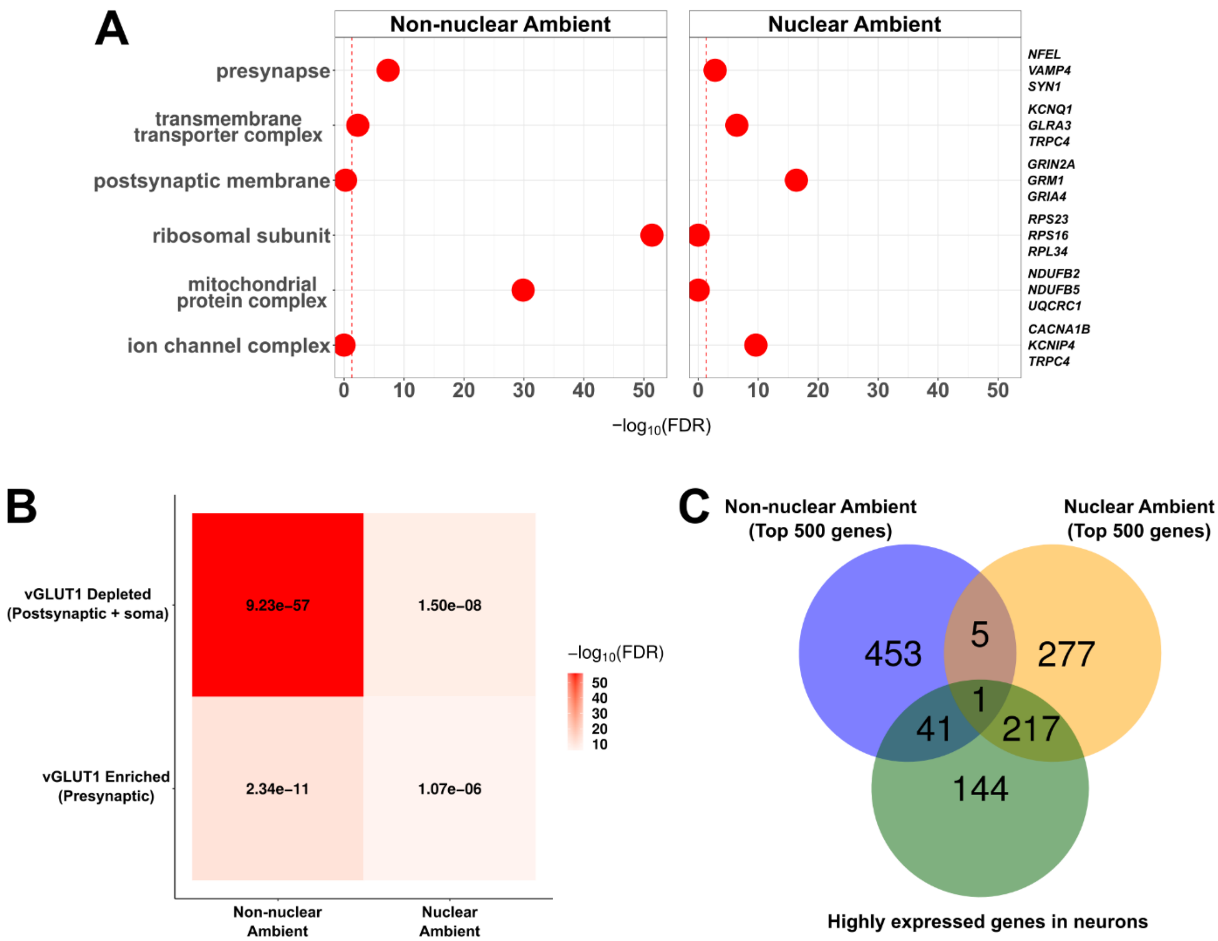


**Figure S2 (Related to Figure 2)**. **Characterization of ambient RNA markers. (A)** Gene ontology (GO) enrichments for non-nuclear and nuclear ambient RNA markers. Example genes per GO term are shown on the right. **(B)** Heatmap of enrichments between presynaptic synaptosomes (vGLUT1 Enriched) and others (vGLUT1 Depleted) (Fisher's exact test; numbers indicate FDR; heatmap color indicates -$\log_{10}$FDR). Genes from the synaptosome dataset were converted from mouse to human symbols prior to enrichment. **(C)** Overlaps between the top 500 most distinct ambient RNA markers and the top 500 highly expressed genes in neurons.

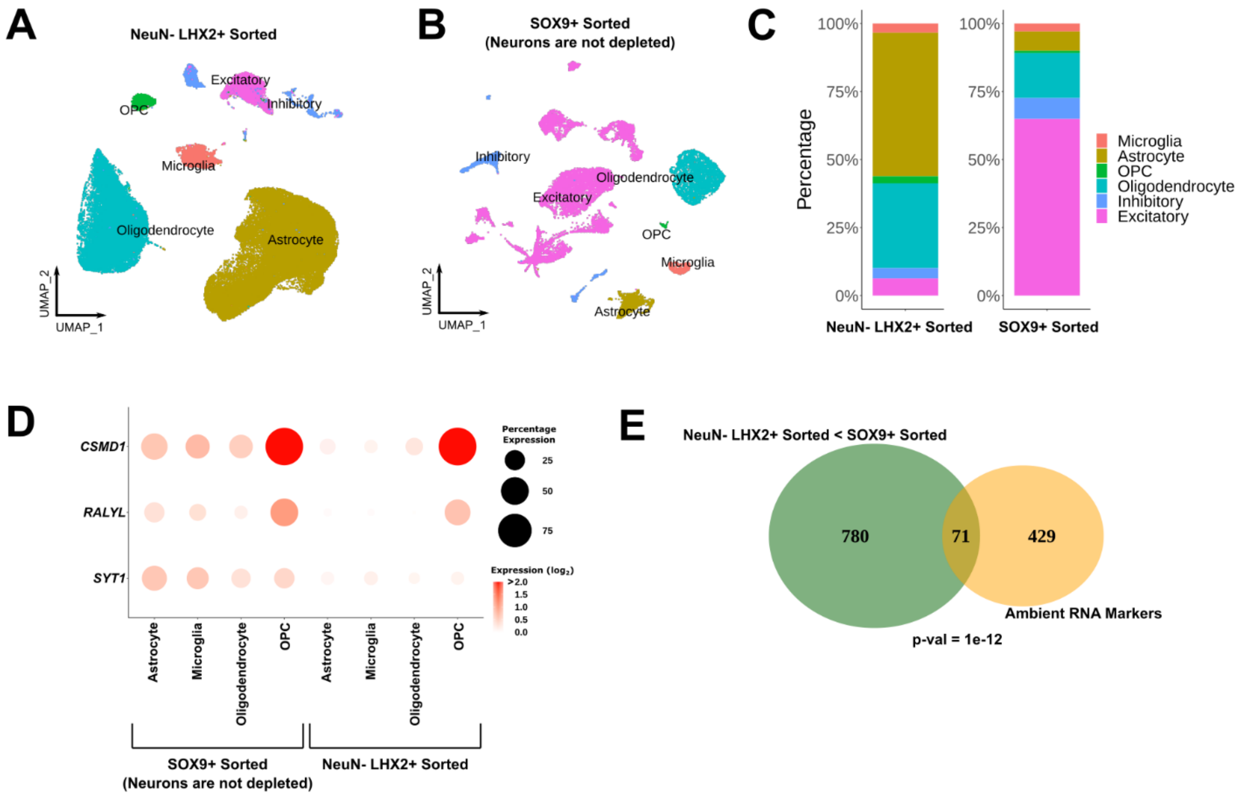


**Figure S3 (Related to Figure 3)**. **Comparison of neuron depleted and non-depleted snRNA-seq datasets from the same study. (A-B)** UMAP plots of sorted datasets with neuron depletion (NeuN- LHX2+ sorted, also used as NeuN- SD3) **(A)** and without neuron depletion (SOX9+ sorted) **(B)**. **(C)** Stacked bar plots of both datasets that show cell type composition by percentage. **(D)** Dot plot of expression levels (normalized, log2 transformed) of selected ambient RNA marker genes across glial cell types. **(E)** Overlap between genes overrepresented in the SOX9+ sorted dataset and the top 500 ambient RNA markers. Only nuclear ambient RNA markers were used since non-nuclear ambient RNAs were removed in the sorted datasets.

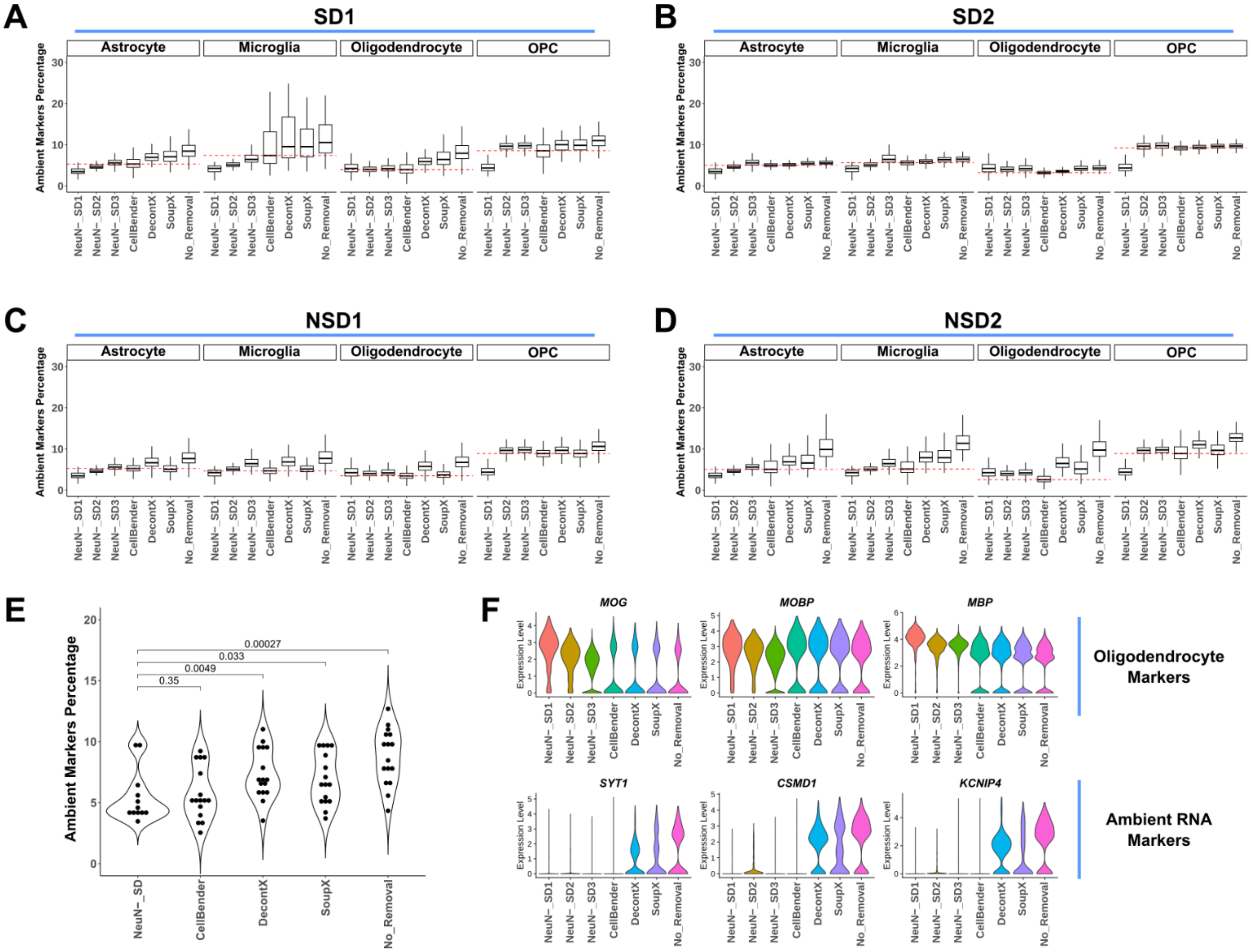


**Figure S4 (Related to Figure 3). Comparison of ambient RNA removal tools.** All tools were evaluated based on the percentage of reads explained by ambient RNA markers in glial cell types. These percentages were then compared to the percentages of ambient RNA markers in NeuN- sorted datasets (NeuN- SDs). **(A-D)** Comparison of ambient RNA marker percentages in SD1 **(A)**, SD2 **(B)**, NSD1 **(C)** and NSD2 **(D)**. Dashed lines in red correspond to the median values of the CellBender result. **(E)** Summary of comparisons in **A-D**. Each dot represents a median value of the boxplots in **A-D**. Numbers indicate p-values from one-sided Wilcoxon rank sum tests between the NeuN- SD results and the results from each ambient RNA removal tool or no removal. **(F)** Violin plots of the expression levels of selected ambient RNAs after implementation of CellBender, DecontX or SoupX. The plot contains the expression values of oligodendrocytes from SD1.

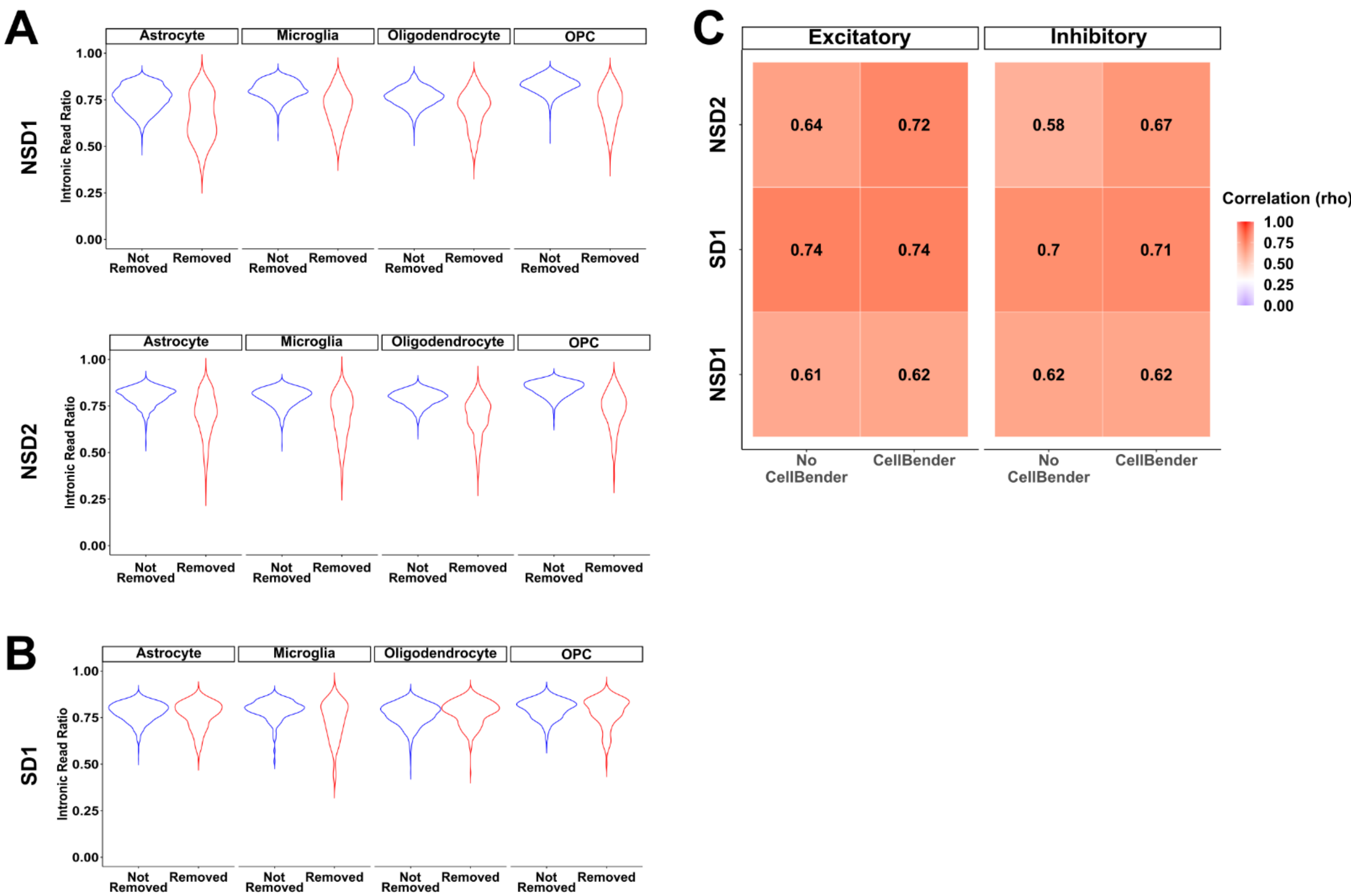


**Figure S5 (Related to Figure 3)**. **Supplementary analyses of CellBender adjustment and subcluster cleaning. (A)** Intronic read ratios per cell barcode in subclusters that were removed due to ambient RNA contamination (red) or not removed (blue) per glial cell type in datasets that did not perform nuclei sorting (NSD1 and NSD2). **(B)** Same as in **(A)** but in a dataset that performed nuclei sorting (SD1). **(C)** Heatmap of Spearman rank correlations of all genes with the NeuN+ sorted dataset (SD2). Correlations were performed per cell type per dataset (y-axis) after each analysis (x-axis). Both numbers and heatmaps indicate the magnitude of correlation coefficient.

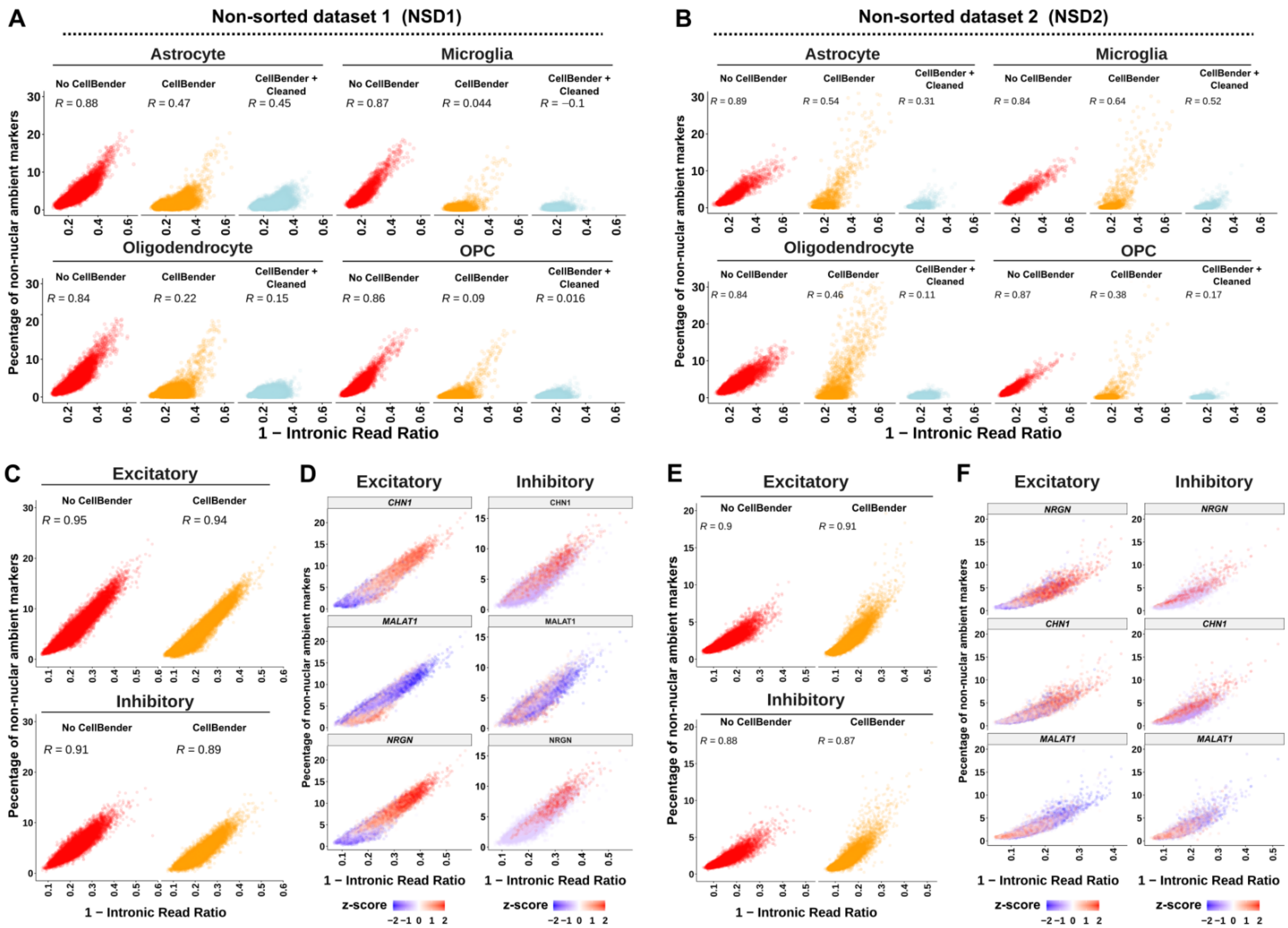


**Figure S6 (Related to Figure 3). Association of non-intronic read ratios and the percentage of non-nuclear ambient markers across nuclei per cell type. (A-B)** Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) in glial cell types either before CellBender (red), after CellBender (orange), or after CellBender + subcluster cleaning (lightblue) using either the NSD1 **(A)** or NSD2 dataset **(B)**. **(C, E)** Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) in excitatory and inhibitory neurons using either the NSD1 **(C)** or NSD2 **(E)** dataset. **(D, F)** Normalized and z-transformed expression levels of non-nuclear ambient markers *NRGN*, *CHN1* and nuclear-retained non-coding gene *MALAT1* in either the NSD1 **(D)** or NSD2 **(F)** dataset. R corresponds to the Spearman's rank correlation coefficient.

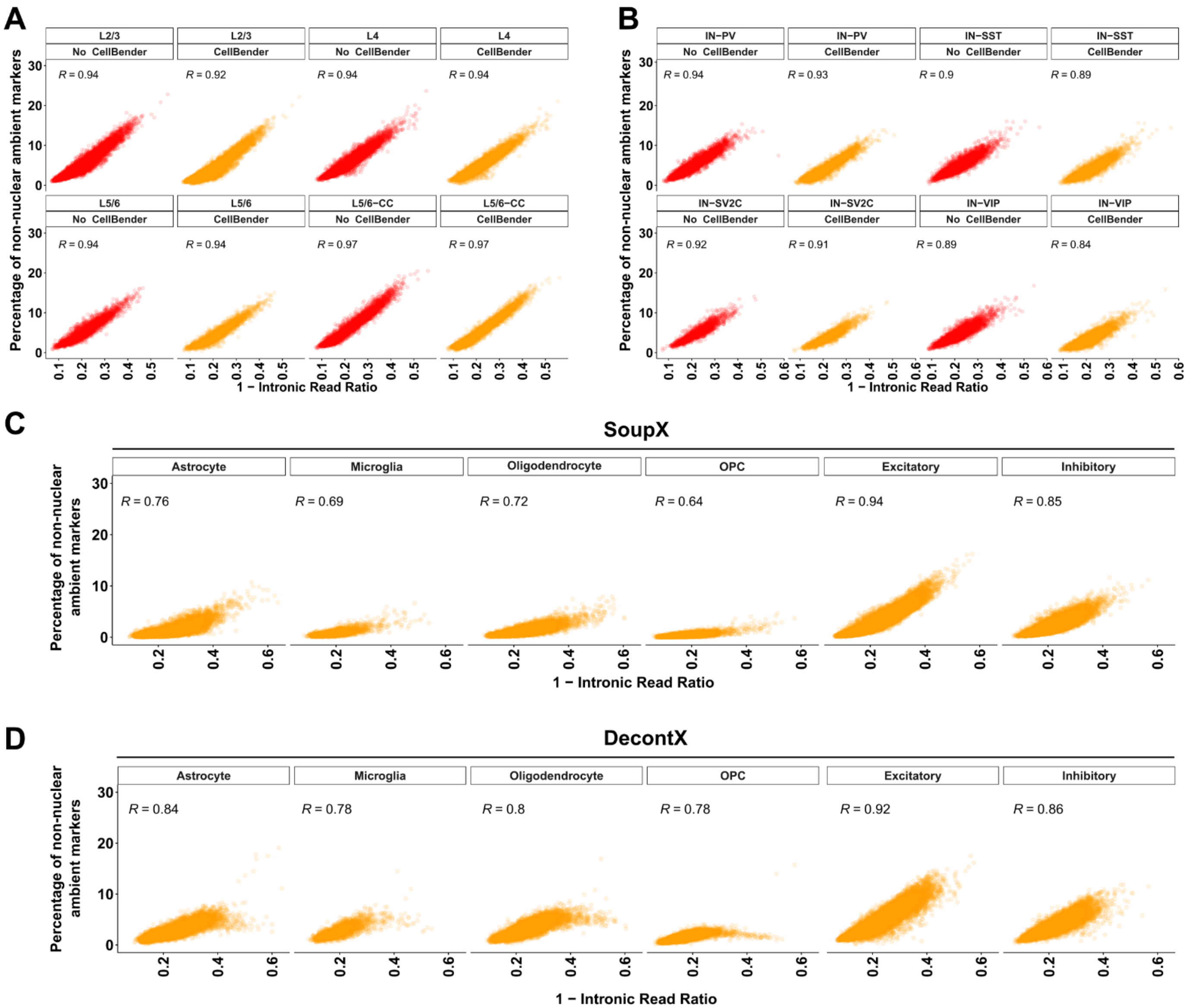


**Figure S7 (Related to Figure 3). Association of non-intronic read ratios and the percentage of non-nuclear ambient markers across nuclei in subtypes or using other tools. (A)** Scatter plots of non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers (y-axis) per annotated excitatory subtype before CellBender (red) or after CellBender (orange). **(B)** Same as **A**, but for inhibitory subtypes. **(C)** Scatter plots of the non-intronic read ratios and percentage of non-nuclear ambient markers per annotated broad cell type after SoupX. **(D)** Scatter plots of the non-intronic read ratios (x-axis) versus the percentage of non-nuclear ambient markers per annotated broad cell type after DecontX. All plots are from the NSD1 dataset. R corresponds to the Spearman's rank correlation coefficient.

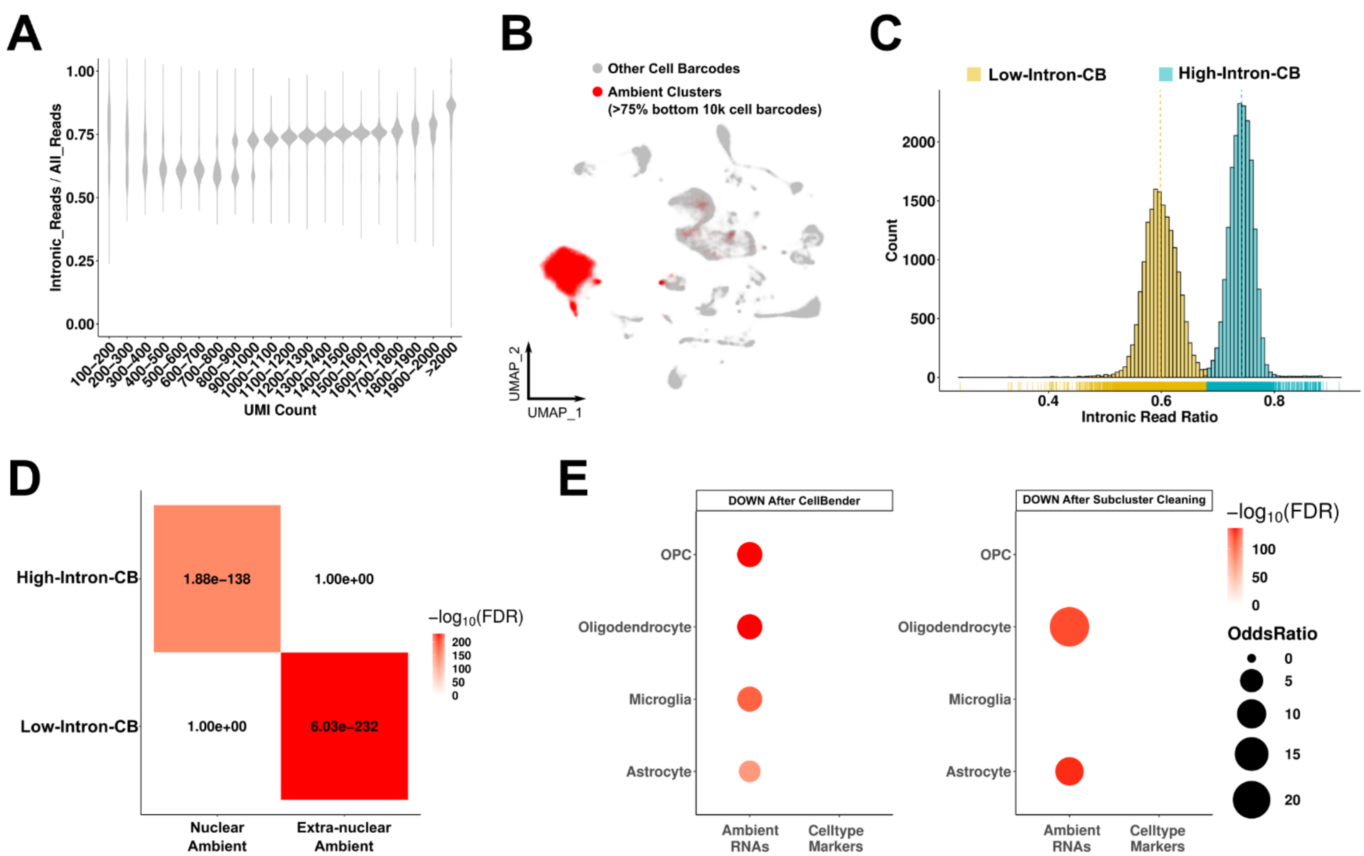


**Figure S8 (Related to Figure 3)**. **Ambient RNAs in a mouse brain snRNA-seq dataset. (A)** The intronic read ratio across increasing UMI counts in a mouse brain snRNA-seq dataset with no nuclei-sorting. UMI counts are divided into intervals of 100 from 100-2000. **(B)** The clusters that contain greater than >75% of filtered cell barcodes are highlighted and named ambient clusters. **(C)** The distribution of intronic read ratios within ambient clusters. Yellow: Low-Intron-CB (CB: Cell Barcodes), blue: High-Intron-CB. **(D)** Heatmap of enrichments between ambient RNA types. Nuclear ambient RNA and non-nuclear ambient RNAs are identified from SD1 and NSD1 (see Figures 1-2). **(E)** Dot plot enrichments between genes with significantly lower expression (DOWN) after CellBender (left) or subclustering steps (right) with ambient RNA markers or cell type markers. Cell types are indicated in the y-axis.

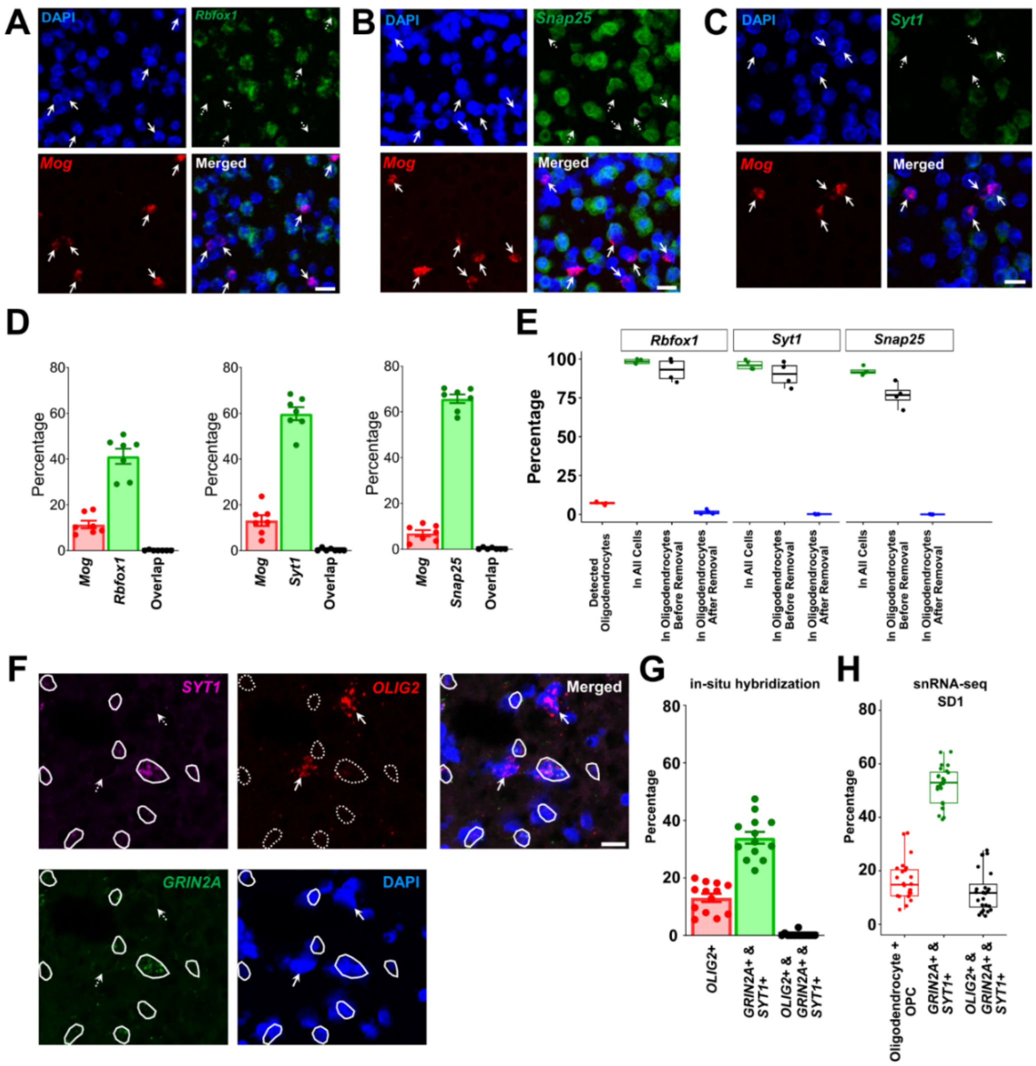


**Figure S9 (Related to Figure 3): In situ hybridization does not detect ambient RNA markers in oligodendrocytes.** (**A-C**) Representative images of smFISH for a marker of mature oligodendrocytes (*Mog*+ cells, solid arrows) and 3 markers of ambient RNAs (*Rbfox1*, *Snap25*, *Syt1*) reveal no overlap (dashed arrows) in adult mouse frontal cortex. (**D**) Quantification of smFISH experiments to indicate the percentage of cells positive for each gene relative to the number of DAPI+ cells (7 sagittal images obtained from 2 mice were quantified for each experiment). Data are represented as mean ± SEM (**E**) The percentage of cells that contain at least one read of each gene per given population for each specified analysis in the adult mouse snRNA-seq dataset. The dataset is the one from Supplementary Figure 8. (**F**) Representative images of smFISH for a marker of mature oligodendrocytes (*OLIG2*+ cells, arrows) or genes that mark neurons (*SYT1* and *GRIN2A*, positive cells are highlighted in circles) in adult human posterior cingulate cortex. (**G**) Quantification of smFISH experiments to indicate the percentage of cells positive for each gene relative to the number of DAPI+ cells (13 images from 6 tissue sections obtained from a total of 3 individuals were quantified for each experiment). Data are represented as mean ± SEM. (**H**) Percentage of cells that were annotated as belonging to the oligodendrocyte lineage or containing at least one read from the given gene in the original SD1 dataset.

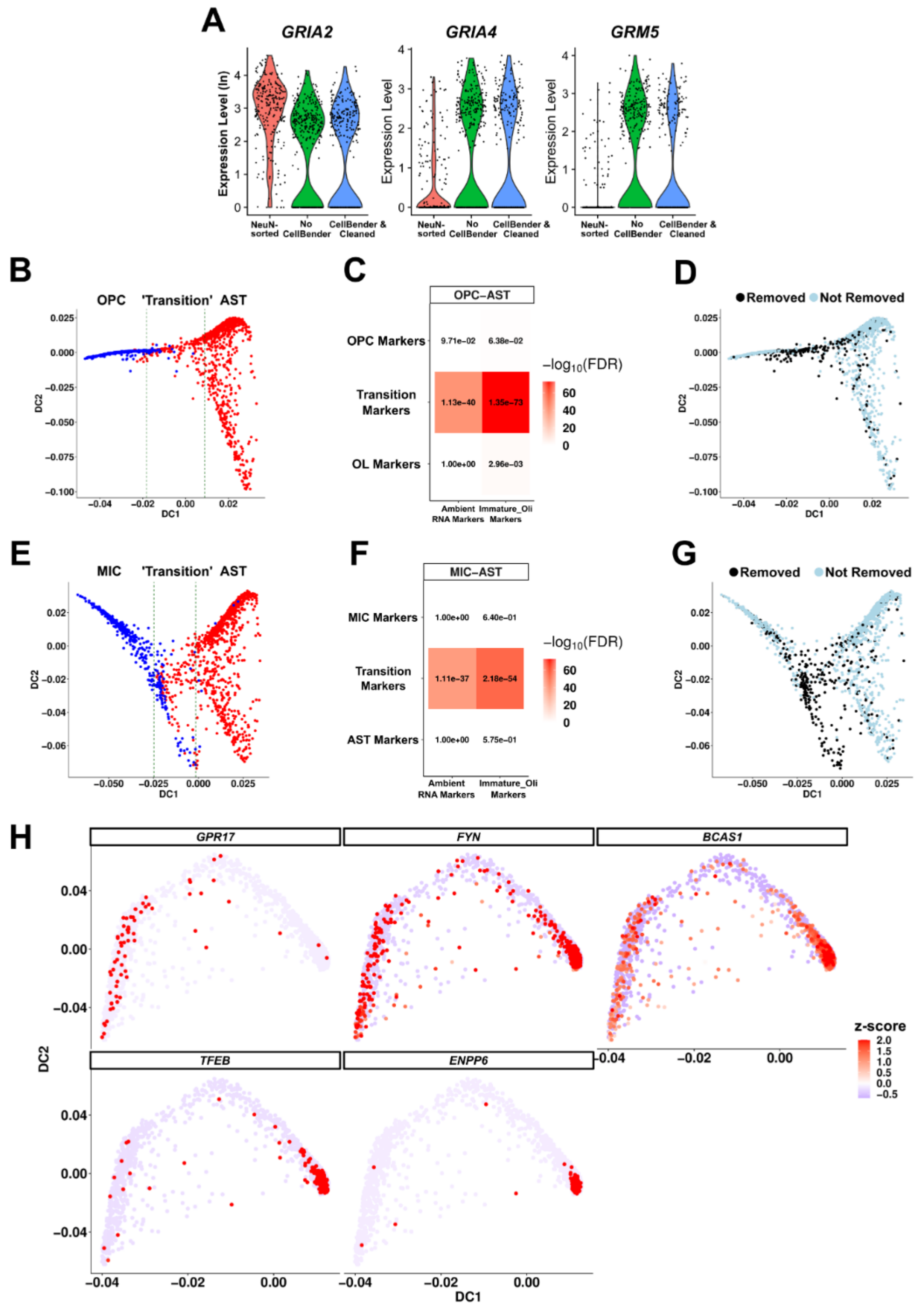


**Figure S10 (Related to Figure 4)**. **Immature oligodendrocytes are explained by ambient RNA contamination. (A)** Violin plots of expression of glutamatergic receptors in NeuN- sorted OPCs and SD1 OPCs with or without ambient RNA removal. **(B)** Pseudotime trajectory of SD1 as reconstructed with *destiny* between OPCs and AST (astrocytes). The 'transition' zone was defined as the 400 nuclei around the middle nucleus based on DC1. **(C)** Heatmaps of enrichments between trajectory zones (OPC, Transition, AST) and ambient RNA or immature oligodendrocyte markers (Fisher's exact test; numbers indicate FDR; color scale is -$\log_{10}$(FDR)). **(D)** The same lineage trajectory as **(B)** with the nuclei removed after subcluster cleaning highlighted. **(E-G)** The same trajectory approach used in **(B-D)** but instead using MIC (microglia) and AST. **(H)** Z-transformed gene expression of COP or Pre-OL (premyelinating oligodendrocyte) markers in the OPC-OL lineage trajectory.

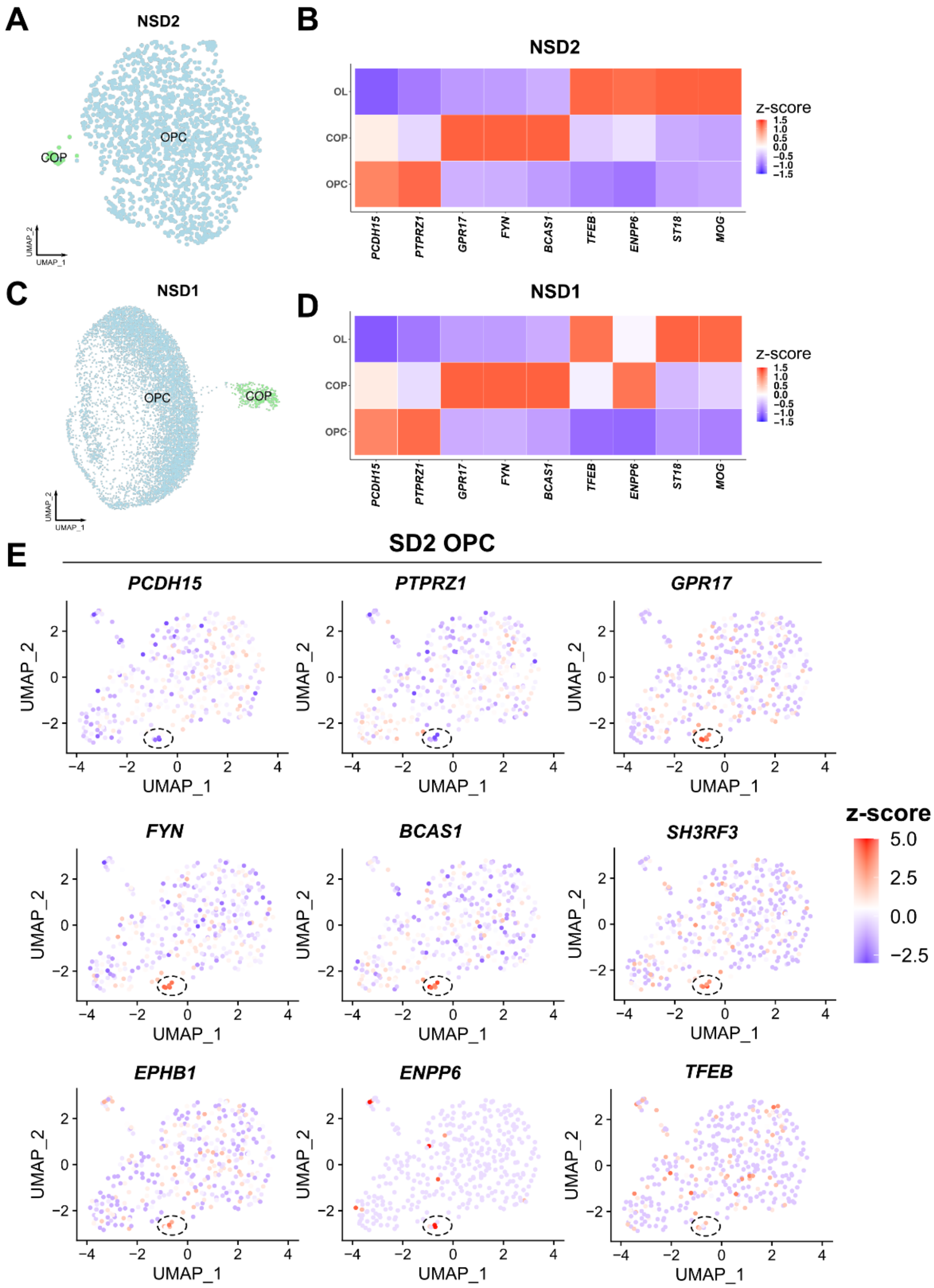


**Figure S11 (Related to Figure 4)**. **Committed progenitor cells in additional adult human brain snRNA-seq datasets. (A, C)** UMAP of OPC subclustering from **(A)** NSD2 or **(C)** NSD1 datasets. COP: committed OPCs. **(B, D)** Heatmap of z-transformed gene expression of oligodendrocyte lineage marker genes (z-scored across cell types per marker gene) from **(B)** NSD2 or **(D)** NSD1 datasets. OPC markers: *PCDH15*, *PTPRZ1*. COP markers: *GPR17*, *FYN*, *BCAS1*. COP and OL markers: *ENPP6*, *TFEB*. **(E)** Feature plots of oligodendrocyte lineage markers genes for OPCs from SD2. The color scheme reflects the z-score per the expression of each gene across cell types (legend on the right side of the plot).

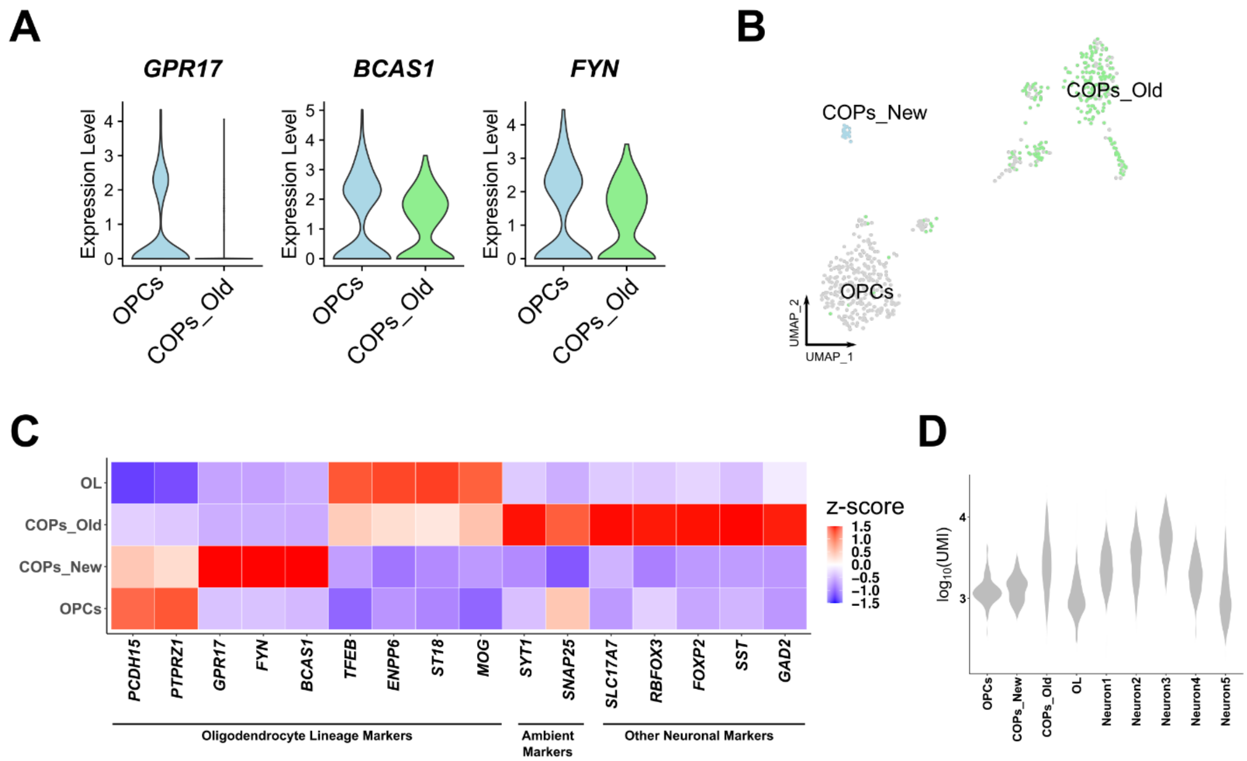


**Figure S12 (Related to Figure 4)**. **Re-assessment of previous COP annotations in human brain snRNA-seq white matter dataset. (A)** Violin plots of expression levels (normalized, log transformed) of COP markers in the original annotation of OPCs and COPs ('COPs_Old') versus expression in nuclei we hypothesize to be true COPs ('COPs_New'). **(B)** UMAP plot of OPCs and COPs. The small subpopulation suspected to be real COPs is indicated as 'COPs_New' whereas the nuclei previously annotated as COPs are shown as 'COPs_Old'. **(C)** Heatmap of z-scored gene expression of oligodendrocyte lineage markers, two top ambient RNA markers (*SYT1*: Nuclear, *SNAP25*: Extra-nuclear) and other neuronal markers in the same dataset. The colors indicate the z-scored expression across cell types per marker gene. Note that z-scored expression only shows the relative expression levels among the four cell type annotations. **(D)** $Log_{10}$ transformed UMI count values per cell type. OL: oligodendrocytes. Neuronal cell types are annotated as in the original publication.