Supplementary Materials for

Single-cell analysis of prenatal and postnatal human cortical development

Dmitry Velmeshev, Yonatan Perez, Zihan Yan, Jonathan E. Valencia, David R. Castaneda-Castellanos, Li Wang, Lucas Schirmer, Simone Mayer, Brittney Wick, Shaohui Wang, Tomasz Jan Nowakowski, Mercedes Paredes, Eric J Huang, Arnold R Kriegstein

Corresponding authors: arnold.kriegstein@ucsf.edu and dmitry.velmeshev@duke.edu

The PDF file includes:

Materials and Methods
Figs. S1 to S7

Other Supplementary Materials for this manuscript include the following:

Data S1 to S6

Materials and Methods

Sample acquisition and selection

Samples were acquired from three different sources. 1) De-identified second-trimester tissue samples were collected at the Zuckerberg San Francisco General Hospital with previous patient consent in strict observance of the legal and institutional ethical regulations. Protocols were approved by the Human Gamete, Embryo, and Stem Cell Research Committee (institutional review board) at the University of California, San Francisco. These fresh tissue samples were dissected and snap-frozen in isopentane on dry ice. 2) De-identified second-trimester, third trimester and early postnatal tissue samples were obtained at the UCSF Pediatric Neuropathology Research Laboratory led by Dr. Eric Huang. These samples were acquired with patient consent in strict observance of the legal and institutional ethical regulations and in accordance to research protocols approved by the UCSF IRB committee. These samples were dissected and snap-frozen either on a cold plate placed on a slab of dry ice or in isopentane on dry ice. 3) Banked de-identified second-trimester, third trimester, early postnatal and adult tissue samples were obtained from the University of Maryland Brain and Tissue Bank through the NIH NeuroBioBank.

For postnatal ages, samples from individuals with known history of brain disorders or brain trauma were excluded from downstream analyses. For prenatal samples, samples with unusual neuropathology following pathological examination, as well as samples positive for commonly tested chromosomal aberrations, were excluded. Prior to performing nuclei isolation and single-nucleus RNA sequencing, samples were screened for RNA quality by collecting 100um-thick cryosections, isolating total RNA and measuring RNA Integrity Number (RIN) using the Agilent 2100 Bioanalyzer instrument. Only samples with RIN >= 6.5 were included in the study.

Nuclei isolation and generation of single-nucleus RNA-seq data using 10x Genomics platform

40 mg of sectioned brain tissue was homogenized in 5 mL of RNAase-free lysis buffer (0.32M sucrose, 5 mM $CaCl^2$, 3 mM $MgAc^2$, 0.1 mM EDTA, 10 mM Tris-HCl, 1 mM DTT, 0.1% Triton X-100 in DEPC-treated water) using glass dounce homogenizer (Thomas Scientific, Cat # 3431D76) on ice. The homogenate was loaded into a 30 mL thick polycarbonate ultracentrifuge tube (Beckman Coulter, Cat # 355631). 9 mL of sucrose solution (1.8 M sucrose, 3 mM $MgAc^2$, 1 mM DTT, 10 mM Tris-HCl in DEPC-treated water) was added to the bottom of the tube with the homogenate and centrifuged at 107,000 g for 2.5 hours at 4°C. Supernatant was aspirated, and the nuclei containing pellet was incubated in 250 uL of DEPC-treated water-based PBS for 20 min on ice before resuspending the pellet. The nuclear suspension was filtered twice through a 30 um cell strainer. Nuclei were counted using a hemocytometer and diluted to 2,000 nuclei/uL before performing single-nuclei capture on the 10X Genomics Single-Cell 3' system. Usually, the target capture of 3,000 nuclei per sample was used; the 10x capture and library preparation protocol was used without modification. Single-nucleus libraries from individual samples were pooled and sequenced on the NovaSeq 6000 machine (average depth 60,000 reads/nucleus).

snRNA-seq data processing with 10X Genomics CellRanger software and data filtering

For library demultiplexing, fastq file generation and read alignment and UMI quantification, CellRanger software v 1.3.1 was used. CellRanger was used with default parameters, except for using pre-mRNA reference file (ENSEMBL GRCh38) to insure capturing intronic reads originating from pre-mRNA transcripts abundant in the nuclear fraction.

Individual expression matrices containing numbers of Unique molecular identifiers (UMIs) per nucleus per gene were filtered to retain nuclei with at least 400 genes expressed and less than 10% of total UMIs originating from mitochondrial and ribosomal RNAs. Individual matrices were combined prior to pre-processing and clustering with Seurat.

snRNA-seq dataset integration, dimensionality reduction, UMAP embedding, clustering and cell type identification

All of the following bioinformatics analysis steps are documented in an R script available at https://doi.org/10.5281/zenodo.7245297.

In order to integrate snRNA-seq datasets, we utilized Harmony (*1*) integration using the 10x Genomics chemistry version as the grouping variable. Downstream data preprocessing, normalization, variable feature selection and PCA was performed using the standard Seurat pipeline (*2*). Selection of significant principal components was done using the elbow method. The selected components were used to perform UMAP embedding and clustering using the Louvain method. The identity of specific lineages and cell types was determined based on expression of known marker genes, as is shown in Figure 1 and Figure S1.

Sex determination

To determine the sex of individuals for which sex information was not available, we aggregated gene expression of all nuclei by individual and plotted individual-wise expression of the following genes: *XIST*, *DDX3Y*, *KDM5D*, *USP9Y*, *ZFY*, *EIF1AY*, *UTY*.

Trajectory reconstruction and isolation of individual lineages

Seurat UMAP coordinates were imported into monocle3 (*3*) for trajectory reconstruction. learn_graph function with custom graph_control options was used to construct the trajectory graph. We noticed that while the original trajectory graph generated by monocle3 corresponded to the major cell lineages, it failed to connect some nodes that passed through populations of cells expressing shared lineage markers. Moreover, some trajectory branches did not correspond to biologically interpretable lineage progression, specifically the branches connecting two mature neuronal cell types containing only adult cells. We corrected these issues by modifying the trajectory according to the following principles: 1) if two terminal nodes failed to be connected but were passing through populations of cells expressing known lineage-specific markers (such as *RORB* for layer 4, *TLE4*/*SEMA4A* for layer 6b, *CUX2* for layer 2-3 and *CUX1* for layer 5-6-IT), we connected these nodes 2) if a branch connected nodes located in two mature cell types, we omitted this branch and 3) based on the first two principles, we isolated the shortest path between the node in the neural progenitor/radial glia cluster and the node in the mature cell type cluster.

Identification of lineage-specific dynamically expressed genes

First, we selected trajectory branches corresponding to specific lineages, as well as the cells along the branches. For the interneuron trajectory analysis, we only selected MGE or CGE cells from the GE progenitors cluster to analyze MGE and CGE-derived INs, respectively. Then, monocle3's Moran's test (graph_test function) was used to identify genes that are dynamically expressed in each lineage. We modified graph_test function to utilize Moran's test with covariates to ensure that our results are not affected by uneven contribution of cells from male and female subjects, different brain regions, as well as cells postmortem interval and 10x chemistry. We selected genes with adjusted p value < 0.05 as statistically significant dynamically expressed genes. To identify lineage-specific genes, we first compressed the single-cell expression data along each lineage by using a sliding window along pseudotime and averaging expression of neighboring cells for each gene. We generated 500 meta-cells in each lineage using this approach. Then, we fit the expression of each gene using a generalized linear model and the following formula: expression ~ splines::ns(pseudotime, df=3). Then, we calculated the area under the curve for the smoothed expression/pseudotime plot for each gene in each lineage across intervals of the sliding window. The difference of under the curve between the lineage of interest and all other lineages was used to rank genes according to their lineage specificity. Moran's p value < 0.05 and an expression difference of at least 20% in one section of the sliding window was used to define lineage-specific genes.

Analysis of single-cell ATAC-seq data and snRNA-seq/scATAC-seq integration

Four scATAC-seq datasets were first remapped to the same hg38 genome reference. Then, a minimal non-overlapping consensus peak set was created based on the peaks from all datasets, and ATAC-seq counts were mapped on this set of peaks using Signac (*4*), and the datasets were combined. Then, gene activity matrix for the combined dataset was generated by counting ATAC peaks in the promoter region and the gene body, using the same parameters as used by the Signac package. For mapping scATAC-seq data on the snRNA-seq dataset, we first integrated the two modalities using Seurat's FindTransferAnchors and the canonical correlation analysis (cca). We used the expression and gene activity of genes variable in the snRNA-seq datasets to perform cca and then used the TransferData function to map the scATAC-seq data on the snRNA-seq space followed by Harmony processing to regress the effect of different scATAC-seq and snRNA-seq chemistries. To map scATAC-seq profiles to the UMAP space and clusters we generated using snRNA-seq data, we identified 100 nearest neighbors for each scATAC-seq cell in the combined snRNA-seq/scATAC-seq space and then calculated the UMAP coordinates and cluster membership in the snRNA-seq space. To validate the accuracy of this procedure, we checked for the specificity of gene activity of cell type markers, as well as for age distribution. This integration and mapping procedure was repeated for the three major lineage classes (excitatory neurons, interneurons and macroglial cells).

SCENIC+ analysis

SCENIC+ requires single-cell transcriptomic and scATAC-seq data mapped to the same category (e.g. cluster) and also recommends generating pseudobulk scATAC-seq profiles prior to the analysis. In order to prepare our data for SCENIC+ analysis, we first selected ATAC-seq cells along the lineage trajectories using a sliding window approach and keeping the cells in cell type-specific clusters. Then, we generated 2500 meta-cell pseudobulk ATAC-seq profiles using the sliding window along each trajectory and summing all ATAC counts. We also generated 2500 meta-cells for the corresponding lineage-specific snRNA-seq profiles and restricted the analysis to lineage and branch-specific genes relevant to each lineage. In order to generate pseudo-multiome profiles from separate snRNA-seq and scATAC-seq datasets, we sorted cells into 10 bins based on the pseudotime progression.

These pseudotime bins were also used to identify differentially accessible regions of chromatin and cis-regulatory topics using cisTopic (*5*), which was used with default settings, except for setting the differential features threshold to 25%. After generating pseudo-multiome profiles, we performed SCENIC+ analysis as described in the tutorial. Significant enhancer-transcription factor-gene relationships in each lineage were exported as the final result.

Identification of sex and region-enriched dynamically expressed genes

To identify male and female-enriched genes in each lineage, we selected cells from only males or females within each lineage and first performed Moran's I test separately for male and female data. Then, we compressed the data and calculated area under the curve for male and female gene expression. Genes with Moran's I statistic >= 0.1, adjusted Moran's p value<0.05 and the area under curve difference between male and female expression >= 50 were considered sex-specific in each given lineage.

Gene ontology analysis

We used ShinyGO (*6*) to perform gene ontology analysis using genes expressed in each lineage as the background gene list. In order to reduce redundancy of the identified GO terms, all significant (adjusted p value < 0.05) terms were used as input to Revigo (*7*) in case more than 10 pathways were identified. The value of the resulting gene list of 0.4 was used. The -log10(p value) and fold enrichment for the resulting non-redundant GO processes were reported.

Analysis of enrichment of disease risk genes

We intersected disease risk gene lists with our list of lineage-specific genes, as well as genes enriched in male and female developmental lineages. We calculated hypergeometric p values for each overlap, using genes expressed in each lineage as the background.

Data visualization

Cell type, gene expression and lineage trajectories for each lineage can be visualized at https://pre-postnatal-cortex.cells.ucsc.edu.

MERSCOPE spatial transcriptomics

Sample preparation was performed according to manufacturer's instructions (MERSCOPE Fresh and Fixed Frozen Tissue Sample Preparation User Guide, Doc. number 91600002). Briefly, fresh snap frozen tissue with a high RNA integrity number (RIN>8) were sectioned (10um thick) using a cryostat and mounted on MERSCOPE functional slides. Sections where then fixed and stored at 70% ethanol for up to two weeks. Sections went through autofluorescence quenching under UV light for 3 hours using the MERSCOPE Photo-bleacher instrument. A Pre-designed panel mix (140 genes) focused on early emerging excitatory lineage-specific genes based on the single-nuclei analysis were used for probe hybridization. Hybridizations were performed at 37°C for up to 48 hours in a humid environment. Post prob hybridization, sections were fixed using formamide and embedded in gel. After gel embedding, tissue samples were cleared using a clearing mix solution supplemented with proteinase K for 24-48 hours at 37°C until no visible tissue was evident in the gel. After clearing was completed, sections were stained for DAPI and PolyT and fixed with formamide prior to imaging. No additional cell boundary stainings were used. The MERSOPE imaging process was done according to the MERSCOPE Instrument Site Preparation Guide (Doc. Number 91500001). Briefly, an imaging kit was thawed at 37°C for 45 minutes, activated and loaded into the MERSCOPE instrument. The flow chamber was then assembled, fluidics were primed, flow chamber filled with liquid and a low-resolution image was taken. Based on DAPI staining, an ROI was chosen for the full imaging experiment. After imaging was complete, data was processed using MERSCOPE proprietary software. Further analysis, visualization, and integration of spatial data, was done using Seurat v5 (Source: vignettes/spatial_vignette_2.Rmd). Putative neuronal layer localization was predicted from co-localization with referenced markers at relevant developmental stages.

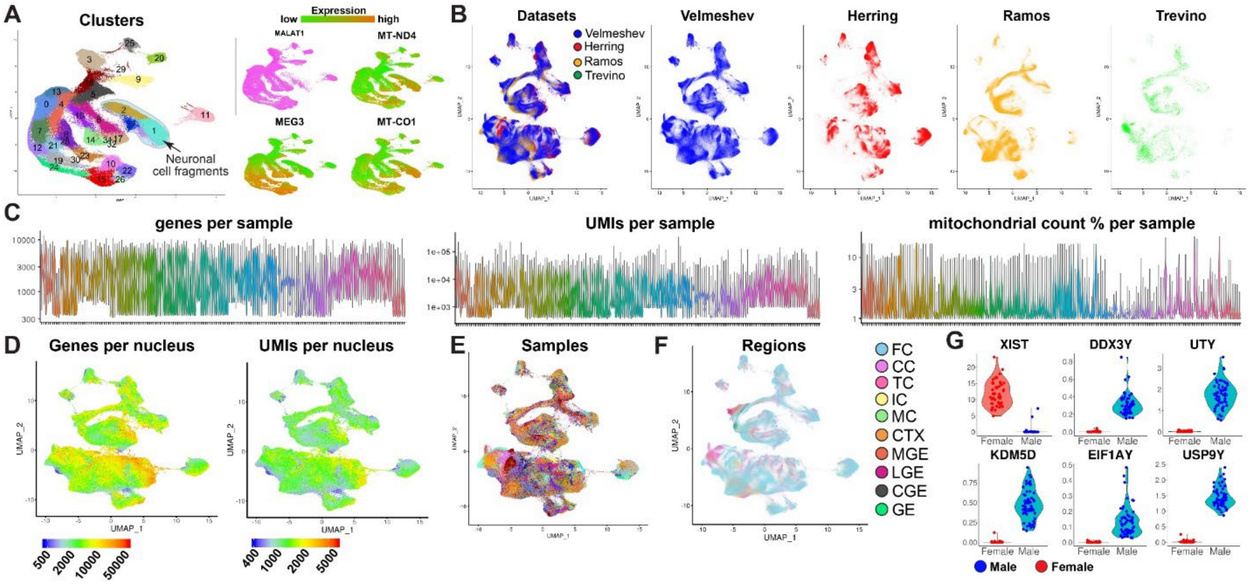

Fig. S1. Technical and biological characteristics of the combined snRNA-seq dataset. A) Identification of the clusters containing neuronal debris. **B)** Integration of the current dataset with previously published datasets. **C)** Gene and UMI counts per nucleus, as well as mitochondrial reads ratio across all samples. **D)** Gene and UMI counts per nucleus across all cell types. **E-F)** Distribution of nuclei from different samples and regions. FC-frontal/prefrontal cortex, CC-cingulate cortex, TC-temporal cortex, IC-insular cortex, MC-motor cortex, CTX-cortex. **G)** Expression of sex-specific genes used to determine sex of samples with unknown status.

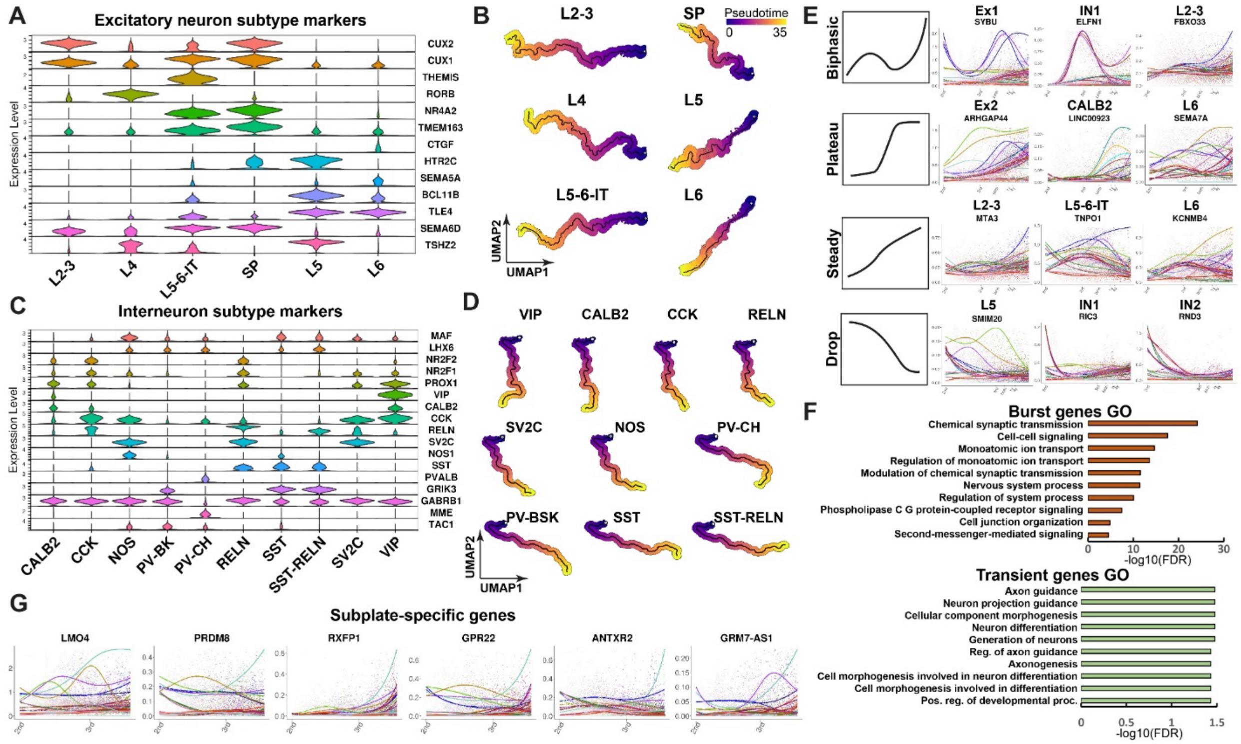

Fig. S2. Excitatory neuron and interneuron lineage analysis. A) Expression of cortical excitatory neuron marker genes used to determine excitatory neuron lineages. **B)** Isolated lineages trajectories for excitatory neuron subtypes. **C)** Markers of interneuron subtypes. **D)** Isolated interneuron trajectories. **E)** Examples of biphasic, plateau, steady and drop expression of lineage and branch-specific genes. **F)** GO pathways enriched for burst and transient neuronal genes. **G)** Top subplate-specific dynamically expressed genes.

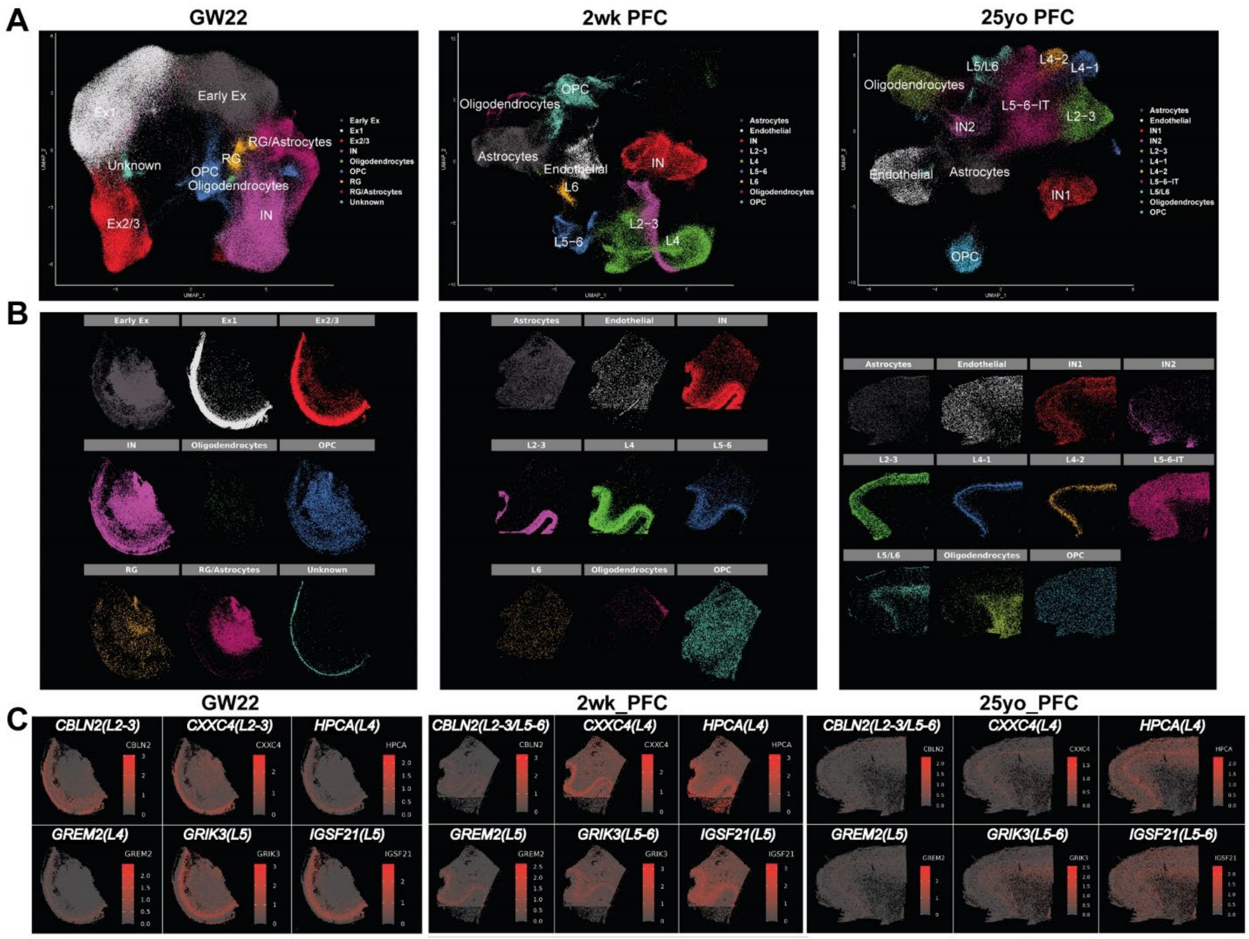

Fig. S3. Spatial transcriptomic analysis of lineage-specific genes across development. **A)** UMAP embedding of annotated clusters. **B)** Spatial localization patterns of individual clusters (cluster colors and spatial location correspond with Fig. 2g). **C)** Spatiotemporal expression of layer-specific markers.

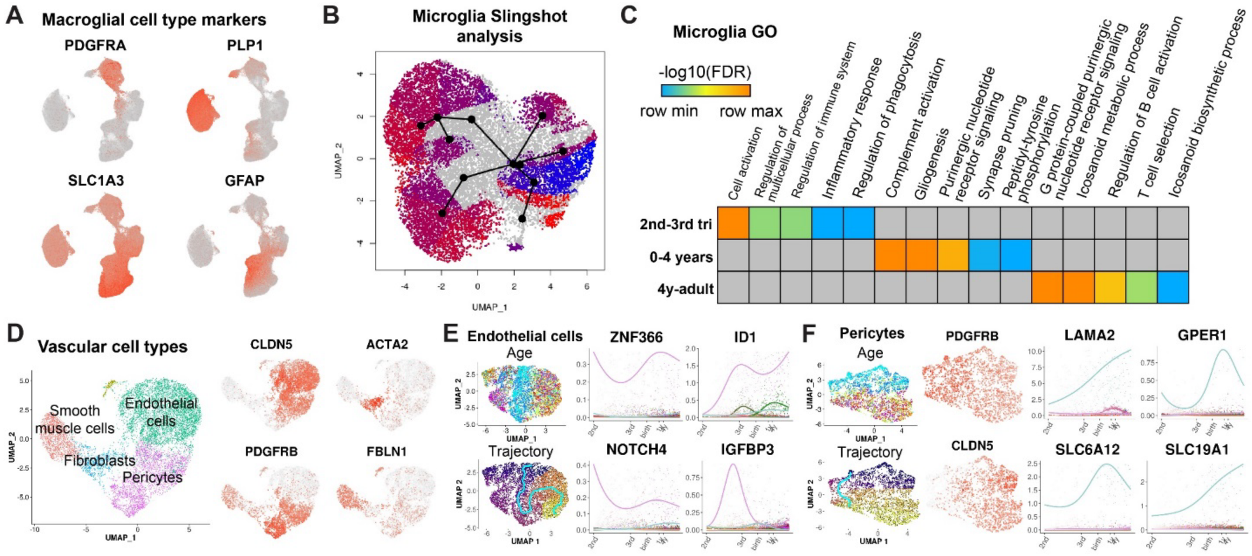

Fig. S4. Analysis of glial and vascular lineages. A) Markers of OPCs, oligodendrocytes, fibrous and protoplasmic astrocytes **B)** Slingshot analysis of microglial lineage trajectories. **C)** Gene ontology analysis developmental microglia genes. **D)** Analysis of vascular cell types. **E-F)** Trajectory analysis of endothelial cells and pericytes.

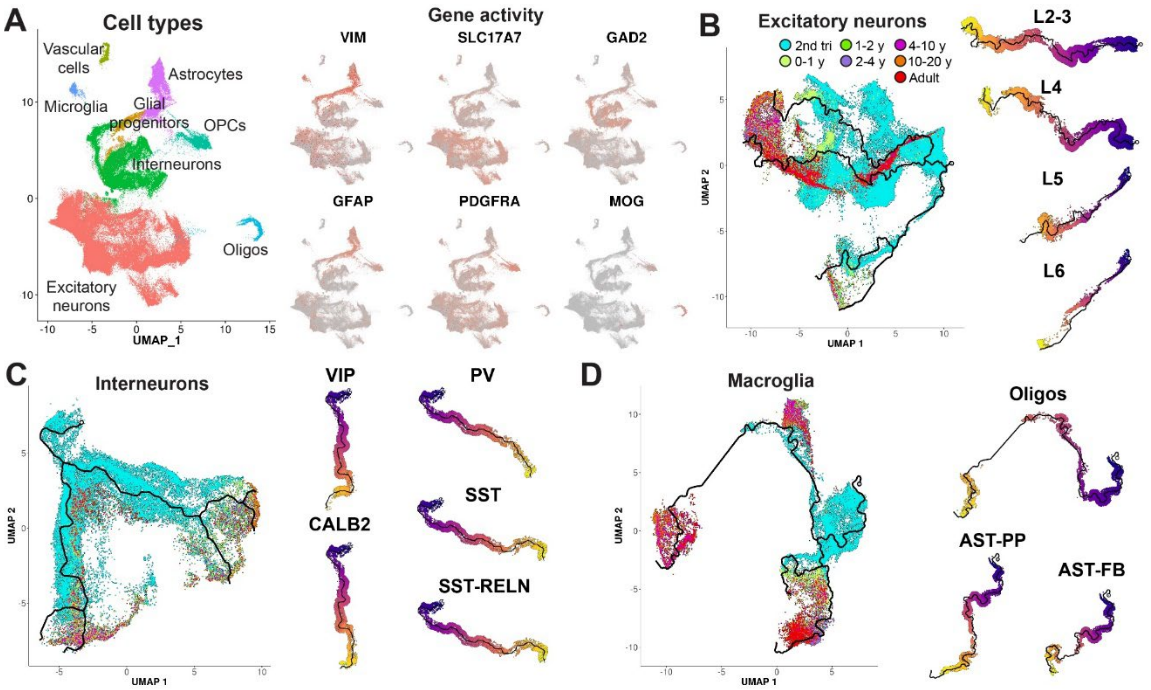

Fig. S5. Mapping developmental scATAC-seq to specific lineage trajectories. A) Gene activities of cell type-specific marker genes. **B-D)** Age distribution and selection of ATAC-seq cells for specific lineages of excitatory neurons **(B)**, interneurons **(C)** and macroglial cells **(D)**.

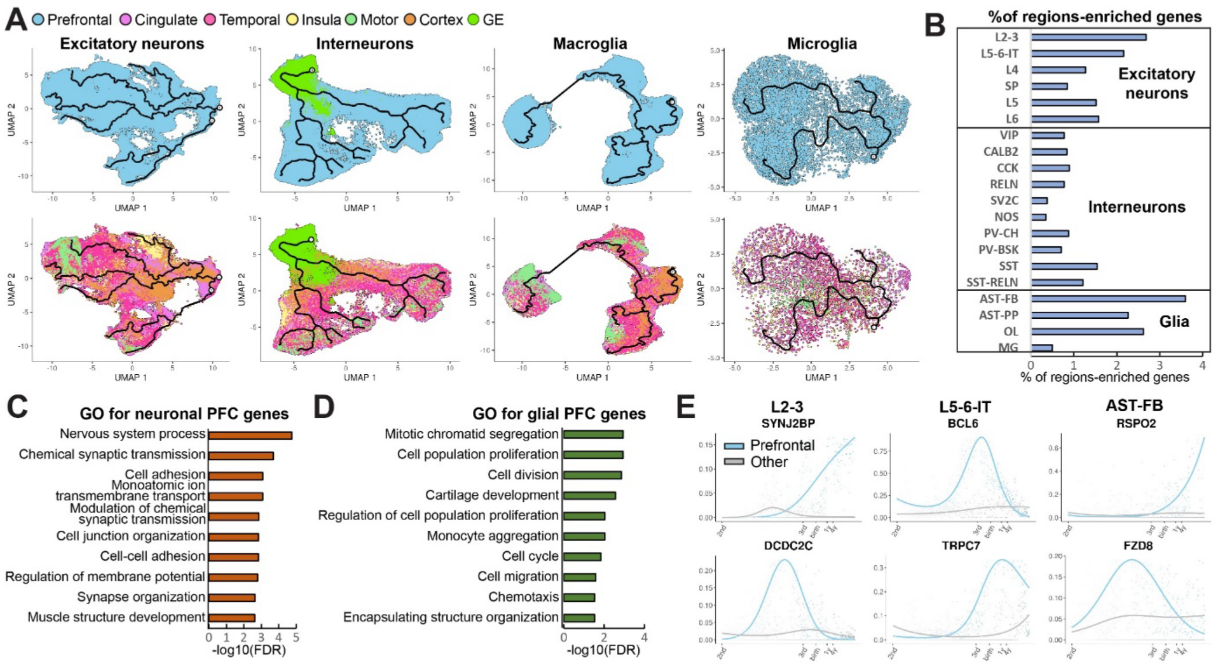

Fig. S6. Frontal cortex-specific developmental programs. A) Cells from the frontal/prefrontal cortex and other cortical regions in the excitatory neuron, interneuron, macroglial and microglial lineages. **B)** Number of PFC-specific genes in neuronal and glial lineages relative to the total number of genes expressed in each lineage. **C-D)** Gene ontology analysis of PFC-specific genes in neuronal and glial lineages. **E)** Examples of top genes enriched in the PFC in specific lineages.

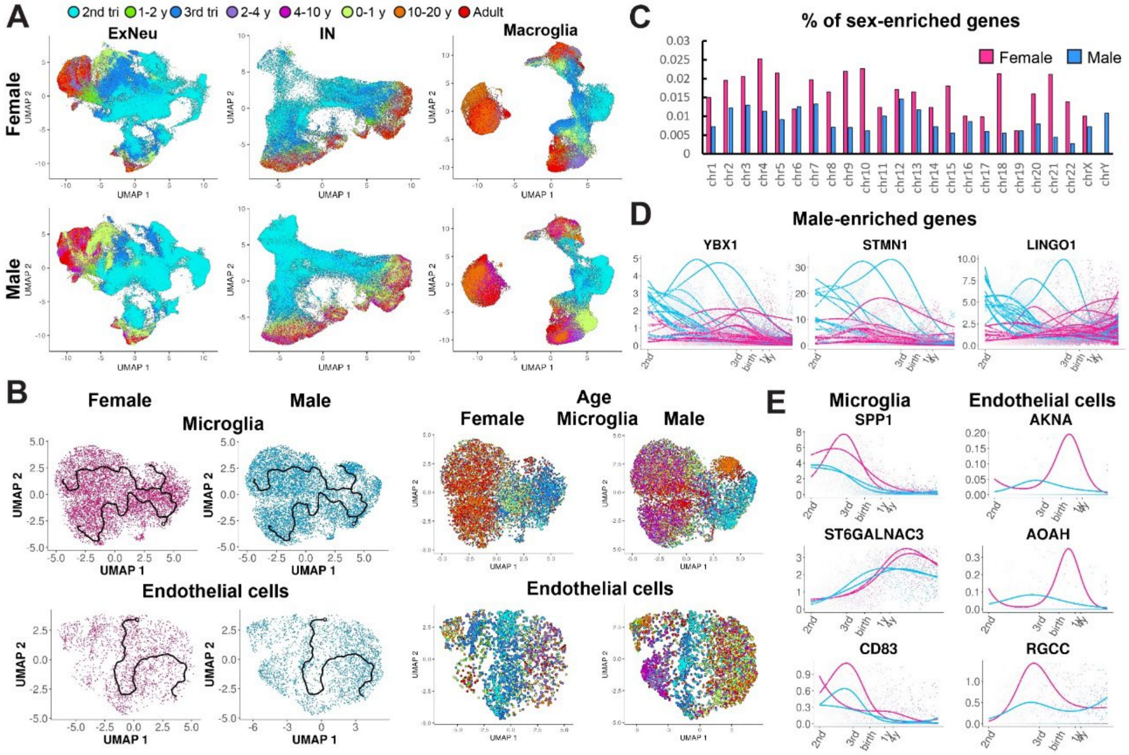

Fig. S7. Analysis of sex and region-enriched genes during microglia and endothelial cell development. A) Female and male microglia and endothelial cell trajectories. **B)** relative number of sex-specific genes per chromosome. **C)** Examples of top male-enriched genes. **D)** Female and male trajectories in microglia and endothelial cells. **E)** Top female-enriched genes expressed in microglia and endothelial cells.

Data S1. Sample and nuclei metadata.

Data S2. Lineage and branch-specific genes.

Data S3. Results of eGRN analysis using SCENIC+.

Data S4. Results of region-specific gene expression analysis.

Data S5. Sex-enriched developmentally regulated genes.

Data S6. Lineage- and sex-specific disease risk genes.

1. I. Korsunsky *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).
2. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
3. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979 (2017).
4. T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, R. Satija, Single-cell chromatin state analysis with Signac. *Nature Methods* **18**, 1333-1341 (2021).
5. C. Bravo González-Blas *et al.*, cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods* **16**, 397-400 (2019).
6. S. X. Ge, D. Jung, R. Yao, ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628-2629 (2019).
7. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).