

A temporal cortex cell atlas highlights gene expression dynamics during human brain maturation

In the format provided by the
authors and unedited

Supplementary Notes

Extended Methods

Nuclei isolation for snRNA-seq

Frozen brain tissue was homogenised in a dounce-homogeniser containing 2 ml ice-cold lysis solution (Nuclei EZ Lysis Buffer [Sigma-Aldrich, NUC101] or Nuclei PURE Lysis buffer [Sigma-Aldrich, NUC201] with 1 mM dithiothreitol [DTT, Promega, P1171, US] and 0.1% Triton X-100 [Sigma-Aldrich, NUC201-1KT, US]). Homogenisation was done 20 times with the loose pestle A followed by 20 times with the tight pestle B. An additional 2 ml lysis solution was added, and the sample was incubated for 5 minutes on ice. The sample was centrifuged at 500 x g for 5 minutes at 4°C after which the supernatant was discarded and the nuclei resuspended in 3 ml ice cold nuclei suspension buffer (1x phosphate-buffered saline [PBS, Sigma-Aldrich, P4417-50TAB, US]), 0.01% bovine serum albumin [BSA, Sigma-Aldrich, A2153-10G, US], and 0.2 U/μl RNasin Plus RNase inhibitor [Promega, N2615, US]). Resuspended nuclei were passed through a 40 μm filter and centrifuged at 900 x g for 10 minutes at 4°C. The supernatant was discarded and pelleted nuclei were resuspended in 3 ml blocking buffer (1xPBS [Sigma-Aldrich, P4417-50TAB, US], 1% BSA [Sigma-Aldrich, A2153-10G, US], 0.2 U/μl RNasin Plus RNase inhibitor [Promega, N2615, US]). To remove myelin debris, 30 μl of myelin removal beads [Milttenyi Biotec. 130-096-733, US] was added to the solution which was mixed by gently pipetting 5 times. The sample was incubated for 15 minutes at 4°C after which it was mixed with 3 ml blocking buffer and centrifuged at 300 x g for 5 minutes at 4°C. The supernatant was removed and the nuclei were resuspended in 2 ml clean blocking buffer. The sample was transferred to a 2 ml tube and placed on a Dynamag magnet for 15 minutes at 4°C. The supernatant was transferred to a new tube and stored on ice. An aliquot of trypan blue stained nuclei was counted using a haemocytometer to determine the nuclei concentration and the volume to use in snRNA-seq library preparation.

DESeq2 age-dependent differential gene expression analysis of snRNA-seq datasets

The unnormalized counts were aggregated across all nuclei for each cluster and sample to generate a 'pseudobulk' counts matrix with the counts from technical replicates collapsed to the level of biological replicates. Genes were filtered to only include those expressed in more than 10% of nuclei for a given cell type. Principal component analysis was performed on each cell type separately in order to assess the variation between samples and determine which variables were contributing most to inter-sample variation from a set of possible variables. The collapsed counts served as input into DESeq2's `DESeqDataSetFromMatrix` function in which the design formula `~single_cell_chemistry + age_group` was specified to treat the `age_group` (paediatric vs adult) as the variable of interest while the effect of `single_cell_chemistry` (version2 vs version3 chemistry) was regressed out. A hypothesis test was performed using the Wald test. The null hypothesis for each gene was that there is no difference in gene expression between the sample groups (i.e Log2Fold Change = 0). Plots of dispersion estimates for each cell type were used to assess the fit of the DESeq2 model to our data (Supplementary Data 11). A Wald test statistic was determined for each gene together with the associated p-value after which the p-values were adjusted for multiple testing using the Benjamini-Hochberg method. Positive

log2 Fold Changes represent genes which are upregulated in paediatric samples compared to adult samples ($p_{adj} < 0.05$).

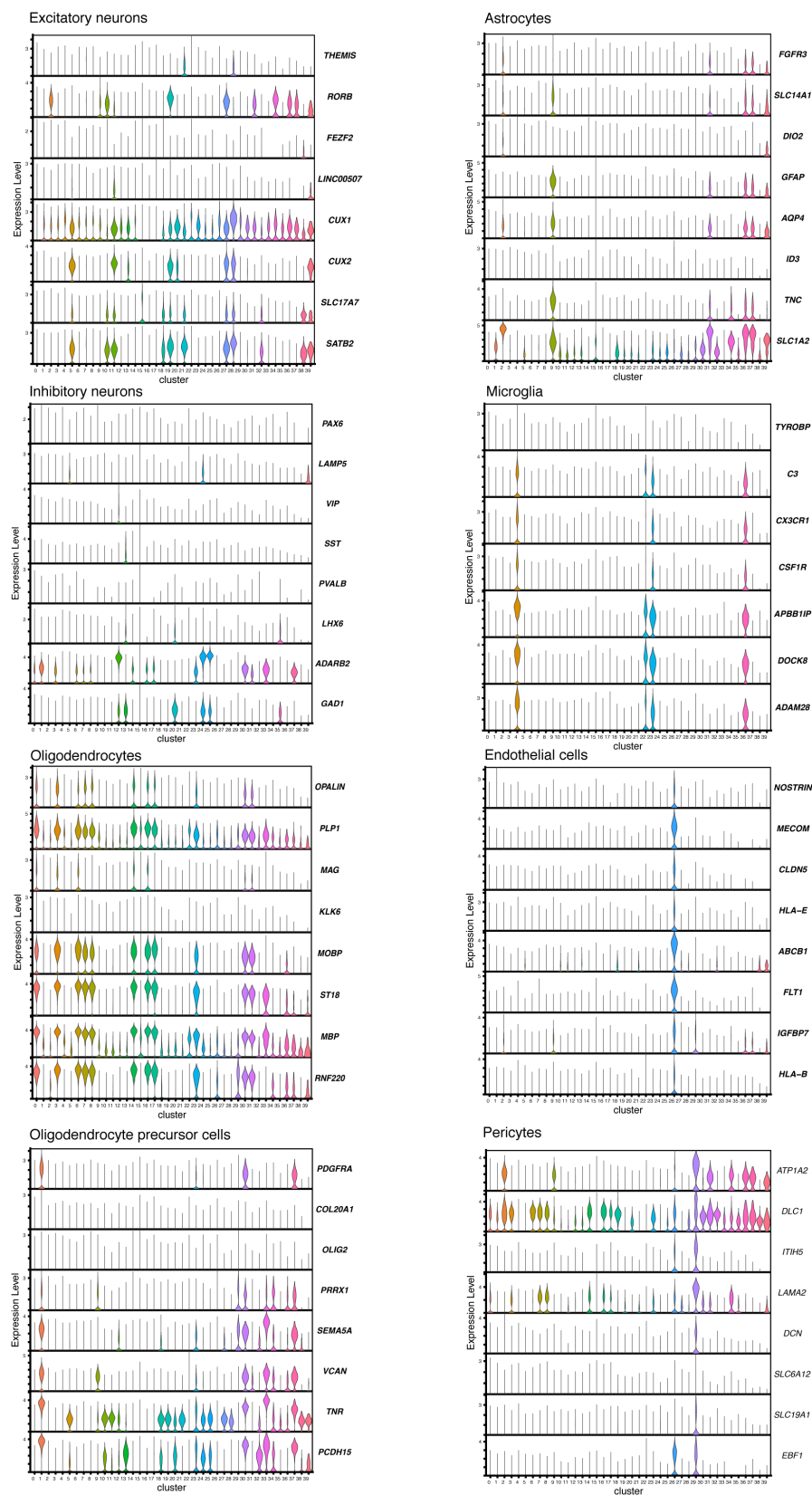
Pseudotime trajectory analysis with psupertime

Psupertime is a supervised approach that uses time-series labels as input to improve the identification of time-varying genes. Each cell type of interest (Exc L3-5 RORB ESR1, Exc L2 LAMP5 LTK, Exc L4-5 RORB FOLH1B, Exc L2-3 LINC00507 FREM3) was individually sub-setted from the Seurat object after which a single cell experiment (sce) object was generated using the log normalized counts (RNA assay, data slot). The donor age (4, 5, 7, 9, 15, 20, 24, 26, 31, 41, 50) was included as metadata in the object. The psupertime function was applied to the sce object with the sel_genes argument specifying all genes to be used. An automatic filtering step was performed to remove genes expressed in fewer than 10% of cells for each cell type. As an output of the function, the beta coefficients for the association of each gene with pseudotime were extracted and plots were generated showing the expression trajectories of the DESeq2 DEGs with pseudotime. Additionally, the overlap between the DESeq2 DEGs and genes changing as a function of pseudotime (Psupertime-relevant genes) was determined.

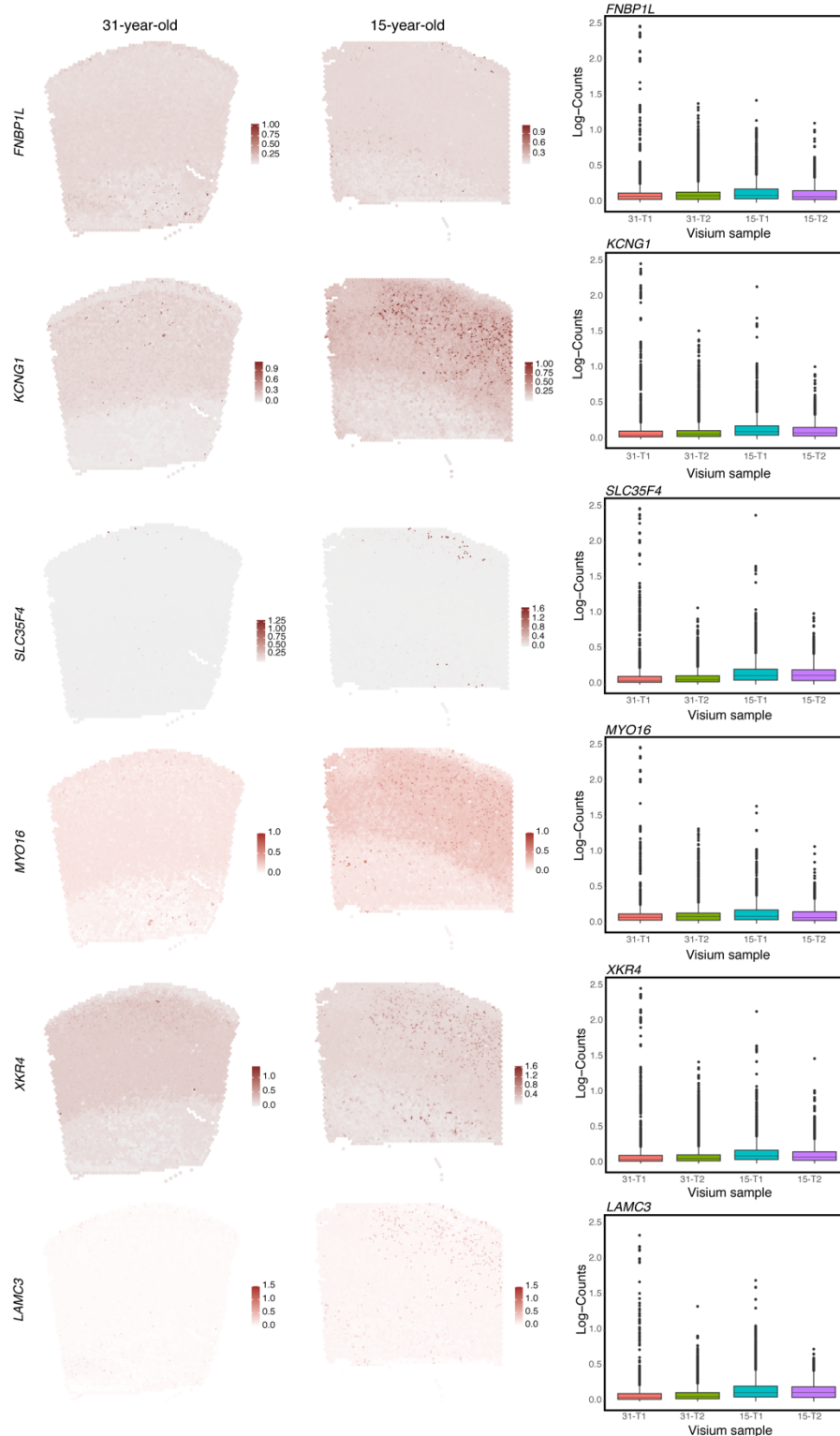
Pathway enrichment analysis of snRNA-seq datasets

GSEA on the DESeq2 output for all genes was performed using the Broad Institute's GSEA software (<https://www.gsea-msigdb.org/gsea/msigdb>) (Code availability: script 20). GSEA aggregates the information from many genes to identify enriched functional pathways, allowing us to interrogate the gene signature changes across all cell types, including those that did not show any significant DEGs⁶⁶. The gene lists for each cell type were queried against the C5 GO Biological Processes collection comprising of gene sets derived from the GO Biological Process ontology. The input lists of genes were ranked according to the $-\log(p\text{-value}) \times \log_2\text{FoldChange}$ for each gene. The parameters specified to the GSEA function included number of permutations (nperm)=1000, minimum gene set size (set_min=15), maximum gene set size (set_max=200), excludes genes that have no gene symbols (collapse)=No_Collapse, value to use for the single identifier that will represent all identifiers for the gene (mode)=Max_probe, normalised enrichment score method (norm)=meandiv, weighted scoring scheme (scoring_scheme)=classic. Positive Normalised Enrichment Scores (NES) represent genes that were upregulated in the paediatric population compared to the adult population ($p < 0.01$ and $q < 0.1$). To visualise the output of universally enriched pathways across multiple cell types, the top 25 most frequently appearing positively and negatively associated terms were plotted. Additionally, for five cell types of interest [which had DEGs meeting the threshold of $p < 0.05$ and $\text{abs}(\log_2\text{FC}) > 0.1$], the top 5 positively associated terms were plotted.

Supplementary Figures 1-6

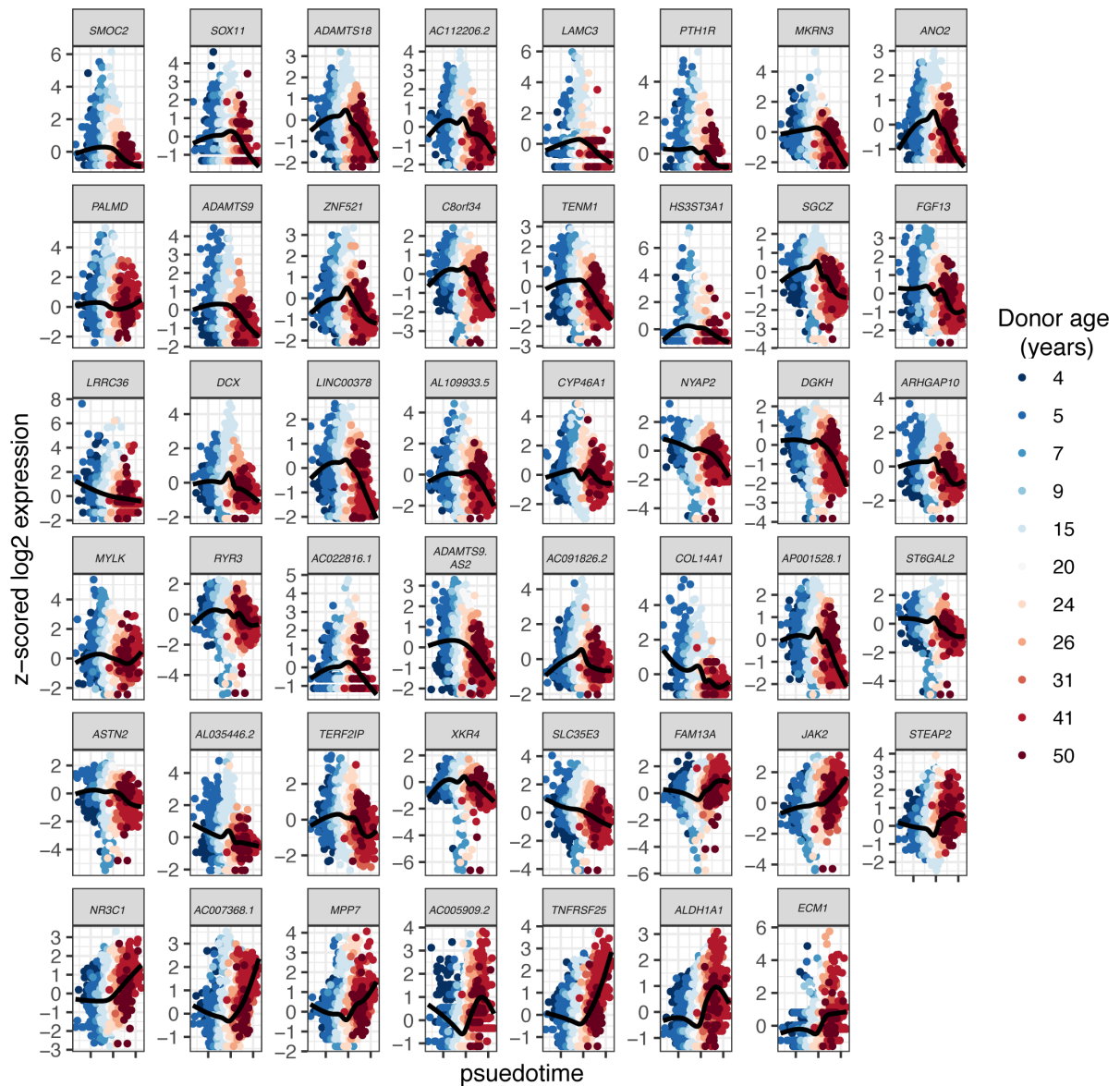


Supplementary Fig. 1: Violin plots showing the expression of known cell type marker genes across the Seurat clusters. These data were used for level 1 annotation of each cluster as one of the indicated major brain cell types (see also Supplementary Data 3).

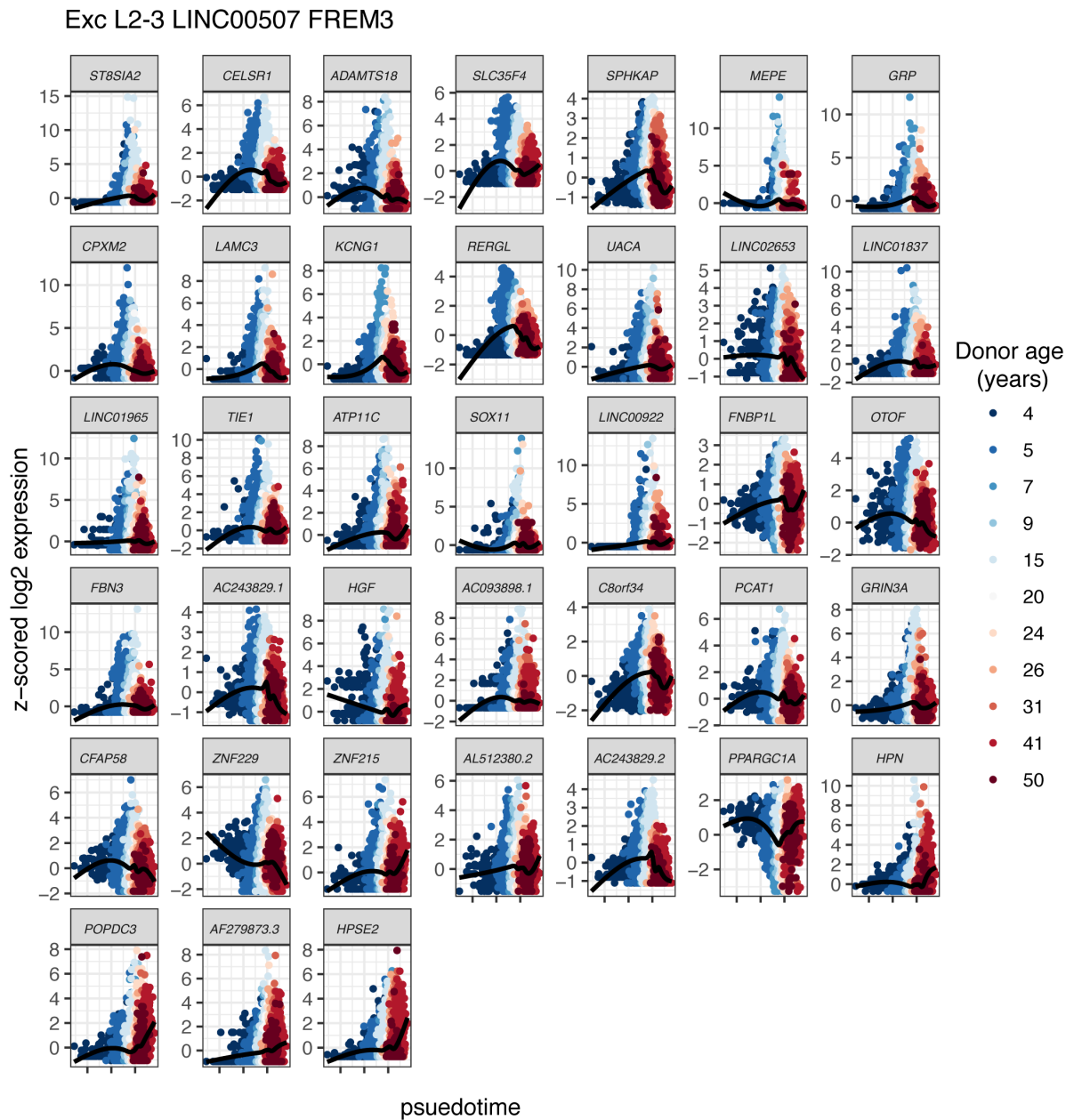


Supplementary Fig. 2: BayesSpace analysis of differentially expressed genes. High resolution Visium spatial gene expression profiles for selected DEGs using BayesSpace analysis to compare sub-spot level expression intensities between 31-year-old and 15-year-old temporal cortex tissue sections. Boxplots show the average gene gene expression (log-counts) across technical replicate (T) samples for the indicated genes for spots with gene expression levels > 0. In all cases, average gene expression is higher in the 15-year-old samples than in the 31-year-old samples. Data in shown in boxplots represents mean $\pm$ sem.

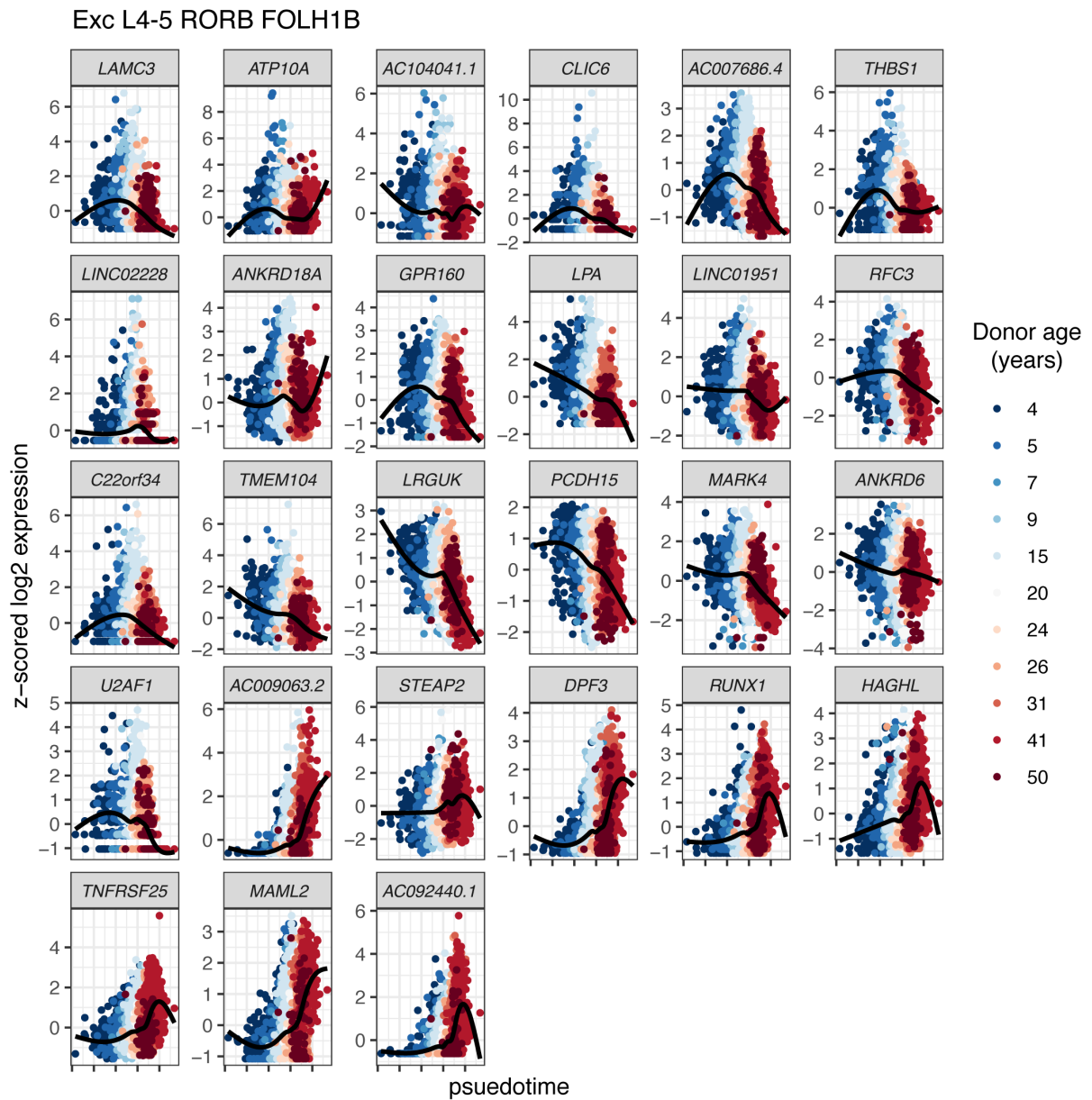
Exc L3-5 RORB ESR1



Supplementary Fig. 3: psupertime gene expression trajectories for all DEGs in Exc_L3-5_RORB_ESR1. The x-axis is the calculated psupertime value for each cell, coloured by sample of origin. The black lines are smoothed curves fit by geom_smooth in the R package ggplot2.

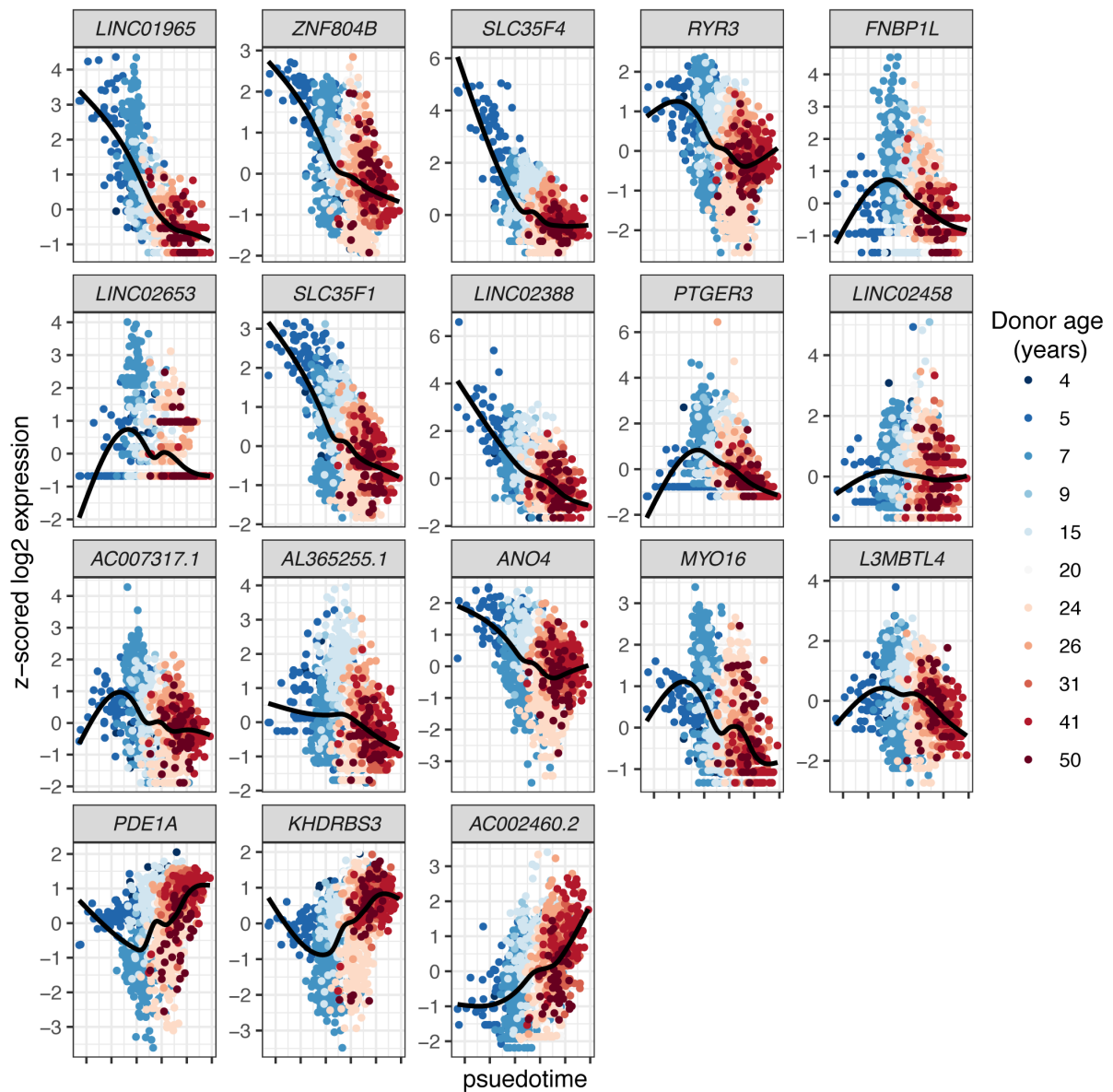


Supplementary Fig. 4: psupertime gene expression trajectories for all DEGs in Exc L2-3_LINC00507_FREM3. The x-axis is the calculated psupertime value for each cell, coloured by sample of origin. The black lines are smoothened curves fit by `geom_smooth` in the R package `ggplot2`.



Supplementary Fig. 5: pspertime gene expression trajectories for all DEGs in Exc_L4-5_RORB_FOLH1B. The x-axis is the calculated pspertime value for each cell, coloured by sample of origin. The black lines are smoothed curves fit by `geom_smooth` in the R package `ggplot2`.

Exc L2 LAMP5 LTK



Supplementary Fig. 6: pspertime gene expression trajectories for all DEGs in Exc_L2_LAMP5_LTK. The x-axis is the calculated pspertime value for each cell, coloured by sample of origin. The black lines are smoothed curves fit by `geom_smooth` in the R package `ggplot2`.

Supplementary Data Legends 1-15

Supplementary Data 1: Summary of snRNA-seq, Visium and MERFISH sample metadata. Samples are ordered by age. The eight “P00” datasets were generated in the Hockman laboratory while the four “Nuc” datasets were generated by Thrupp et al. (2020).

Supplementary Data 2: Summary of average quality control metrics for snRNA-seq datasets across nuclei for each sample before and after filtering. Several measures for quality control were evaluated on a per sample basis including the sequencing saturation, the mean number of reads per nucleus, the number of barcodes, the median number of genes detected per nucleus, the median number of UMIs detected per nucleus, and the number of doublets removed.

Supplementary Data 3: Label transfer annotation of snRNA-seq datasets using the Allen Brain Map MTG atlas as a reference. **Sheet 1**, Manual annotation of clusters into major cell types (level 1 annotation). **Sheet 2**, Number of nuclei per level 1 annotation category per sample after filtering discordant nuclei. **Sheet 3**, Number of nuclei per MTG cell type per sample after filtering discordant nuclei. The number of barcodes corresponding to each MTG cell type and sample is shown. Additionally, the total, minimum, and maximum number of nuclei per cell type and sample was computed. The number of cell types represented per sample was also determined. **Sheet 4**, Number of removed nuclei per level 1 annotation category per sample. **Sheet 5**, Number of removed nuclei per MTG cell type per sample. **Sheet 6**, Subtraction matrix comparing cosine similarity scores (i.e. similarity score for each cell subtype compared to the MTG cell Atlas as in Fig. 1d) for paediatric dataset to the adult dataset. Values are the paediatric scores minus the adult scores. **Sheet 7**, p-values and the test performed for each cell type (Student's two-tailed t-test, two-tailed Brunner-Munzel test or Welch two-tailed t-test) when comparing the proportion of nuclei between male and female samples. **Sheet 8-10**, p-values and the test performed for each cell type (Student's two-tailed t-test, two-tailed Brunner-Munzel test or Welch two-tailed t-test) when comparing the proportion of nuclei (sheet 8), number of genes (sheet 9; see Extended Data Fig. 3b) and number of UMIs (sheet 10; see Extended Data Fig. 3c) for each cell type between paediatric and adult samples shown. All p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

Supplementary Data 4: Summary of average quality control metrics for Visium datasets. Several measures for quality control were evaluated on a per sample basis including the sequencing saturation, the percentage of read mapped to the transcriptome, the number of spots under the tissue, the average number of nuclei per spot determined by Vistoseg analysis, the mean reads detected per spot, the median genes detected per spot, the total number of genes detected, the median UMI Counts per Spot and the total number of nuclei.

Supplementary Data 5: NS-Forest minimal marker analysis. **Sheet 1**, Statistical tests evaluating the expression of Aevermann et al. (2021) markers in our datasets. P-values were determined using Seurat FindMarkers. All p-values were adjusted for multiple testing using the Benjamini-Hochberg method (p_val_adj column) and the Holm method (p_adj_multiple_genes column). **Sheets 2-3**, Metadata for

each feature identified by NS-Forest marker in the down-sampled paediatric (**sheet 2**) and down-sampled adult (**sheet 3**) datasets describing the cell type, the F-beta score for each marker gene, overlap with Aevermann et al. (2021) and Hodge et al. (2019), uniqueness to the age group of interest, coding status, and uniqueness to the associated cell type as shown in Fig. 3. As input to NS-Forest, all datasets (six paediatric and six adult) were randomly down-sampled such that the total number of nuclei per sample was equal to the sample with the fewest number of nuclei.

Supplementary Data 6: gProfiler analysis of NS-forest markers. Sheet1-3, Significantly enriched GO terms associated with the paediatric (**sheet1**), adult (**sheet2**) and paediatric plus adult minimal marker genes identified by NS-forest. (**Sheet3-5**) Significantly enriched GO terms associated with shared (i.e. associated with both adult and paediatric samples) or paediatric-specific marker genes for Oligo L1-6 OPALIN. Terms for which “highlighted” is true are driver terms. gProfiler statistics were generated with g:GOST, which uses Fisher's one-tailed test, and the default g:SCS method for computing multiple testing correction for p-values.

Supplementary Data 7: Summary of metadata for NS-Forest markers with a binary expression score (> 0.7) per cell type across the paediatric and adult datasets. The number of shared markers, the number of markers unique to paediatric samples, and the number of markers unique to adult samples is shown for each cell type. The number of nuclei per cell type is shown for the combined paediatric and adult down-sampled datasets, the down-sampled paediatric datasets, and down-sampled adult datasets.

Supplementary Data 8: Overlap NS-Forest markers with a binary expression score (> 0.7) per cell type between the paediatric and adult datasets. NS-Forest markers with a binary expression score (> 0.7) per cell type were extracted for the down-sampled paediatric and down-sampled adult datasets. Each sheet represents 1 of 75 cortical cell types and the NS-Forest features which were shared (intersect) between the paediatric and adult datasets, unique to paediatric datasets, or unique to adult datasets are shown.

Supplementary Data 9: DESeq2 output of all genes tested for differential expression between paediatric and adult brains per cell type. (Sheet 1-75) Differential expression analysis was performed using DESeq2's two-tailed Wald Test for each cell type separately. Genes were filtered prior to testing to only include those expressed in > 10% of nuclei for that cell type across all paediatric and adult datasets. The associated log₂FoldChanges, p-adjusted values (padj), and description of each feature are shown. Positive log₂FoldChanges represent genes upregulated in paediatrics versus adults. See DESeq2 documentation for explanation of NA values (<https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#why-are-some-p-values-set-to-na>).

Supplementary Data 10: DESeq2 output of significant DEGs only between paediatric and adult brains in a subset of cell types. (sheet 1-21) Significant DEGs (padj < 0.05) for cell types shown in Fig. 5a. Differential expression analysis was performed using DESeq2's two-tailed Wald Test for each cell type

separately. The associated $\log_2$ FoldChanges, p-adjusted values (padj), description, percentage of paediatric nuclei expressing the gene, percentage of adult nuclei expressing the gene, average normalised expression across paediatric nuclei, and average normalised expression across adult nuclei are shown. The difference in the paediatric and adult values for percentage of nuclei and average normalised expression is also shown. Positive $\log_2$ FoldChanges represent genes upregulated in paediatric versus adults datasets. See DESeq2 documentation for explanation of NA values (<https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#why-are-some-p-values-set-to-na>). (sheet 22-24) EA (sheet 6), IQ (sheet7) and HAR (sheet8) associated DEGs and their associated cell types.

Supplementary Data 11: Dispersion plots for DESeq2 analysis. Plots of dispersion estimates were generated for each cell type to assess the fit of the DESeq2 model to the data.

Supplementary Data 12: DEG expression compared across samples at the level of single nuclei. Violin plots for the all DEGs for cell types with >6 DEGs identified by DESeq2 analysis. Black data points represent the normalised expression level for each analysed nucleus in each sample. Blue dots indicate the mean expression level.

Supplementary Data 13: Average DEG expression compared across paediatric and adult samples. Violin plots for the all DEGs for cell types with >6 DEGs identified by DESeq2 analysis. Black data points represent the average normalised exp

Supplementary Data 14: psupertime coefficients. The calculated psupertime coefficients for each gene for indicated excitatory neuron subtypes that showed the highest number of DEGs. Genes with non-zero psupertime coefficients represent genes that are relevant to the ordering of the cells in pseudotime.

Supplementary Data 15: GSEA terms associated with each cell type showing enriched or depleted pathways in paediatric versus adult samples. GSEA was performed using DESeq2's output gene lists for each cell type ranked according to the $\log_2$ FoldChange* $-\log_2$ (padj) for each gene. All DESeq2-tested genes served as input into GSEA (genes were expressed in > 10% of nuclei for the cell type of interest). Matrix shows the corresponding positive (sheet 1) and negative (sheet 2) NES values for each GSEA term (y axis) and cell type (x axis) based on the analysis using the ranked list of genes for each cell type. Terms were filtered to only include significantly associated terms ($p < 0.01$, $q < 0.1$). Positive NES values indicate pathways that are enriched in paediatric versus adult samples; negative NES values indicate pathways that are depleted in paediatric versus adult samples. The total number of terms per cell type and the total number of cell types associated with a given term are shown.