Title: Cell type-specific gene expression dynamics during human brain maturation

Christina Steyn[1,2], Ruvimbo Mishi[1,2], Stephanie Fillmore[1,2], Matthijs B. Verhoog[1,2], Jessica More[1,2], Ursula K. Rohlwink[2,3], Roger Melvill[3], James Butler[2,4], Johannes M. N. Enslin[2,3], Muazzam Jacobs[2,5,6,7], Sadi Quiñones[8,9], Chris G. Dulla[8], Joseph V. Raimondo[1,2,5], Anthony Figaji[2,3] and Dorit Hockman[1,2,*]

Affiliations:

[1]Division of Cell Biology, Department of Human Biology, University of Cape Town, Cape Town, South Africa

[2]Neuroscience Institute, University of Cape Town, Cape Town, South Africa

[3]Division of Neurosurgery, Department of Surgery, University of Cape Town, Cape Town, South Africa

[4]Division of Neurology, Department of Medicine, University of Cape Town, Cape Town, South Africa

[5]Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[6]Division of Immunology, Department of Pathology University of Cape Town

[7]National Health Laboratory Service, South Africa

[8]Department of Neuroscience, Graduate School of Biomedical Sciences, Tufts University School of Medicine, Boston, MA, USA

[9]Graduate School of Biomedical Science, Tufts University School of Medicine, Boston, MA, USA

*Corresponding author

Abstract

The human brain undergoes protracted post-natal maturation, guided by dynamic changes in gene expression. To date, studies exploring these processes have used bulk tissue analyses, which mask cell type-specific gene expression dynamics. Here, using single nucleus (sn)RNA-seq on temporal lobe tissue, including samples of African ancestry, we build a joint paediatric and adult atlas of 54 cell subtypes, which we verify with spatial transcriptomics. We explore the differences in cell states between paediatric and adult cell types, revealing the genes and pathways that change during brain maturation. Our results highlight excitatory neuron subtypes, including the LTK and FREM subtypes, that show elevated expression of genes associated with cognition and synaptic plasticity in paediatric tissue. The new resources we present here improve our understanding of the brain during a critical period of its development and contribute to global efforts to build an inclusive cell map of the brain.

Introduction

The adult human brain is a complex assembly of diverse cell types, which has now been defined with unprecedented accuracy using single cell transcriptomics[1-4]. This adult transcriptomic signature is set up over a protracted period of development, which begins in the embryo and continues after birth as the brain matures. However, while the single cell diversity of the embryonic human brain has been explored[5,6], little is known about how these cell type-specific gene expression profiles change during childhood. Existing studies of gene expression dynamics during human brain development and maturation have used bulk transcriptomic approaches, which revealed a dramatic period of global gene expression change during the late foetal/early infancy transition, that stabilises during childhood and adolescence[6-10]. Bulk transcriptomics, however, cannot reveal the more subtle, cell type-specific changes in gene expression that drive brain maturation from childhood, through adolescence to adulthood.

Childhood and adolescence are periods of important changes in brain structure, during which neuronal connections are refined and strengthened. While synaptogenesis peaks in the early postnatal period, synaptic pruning activity begins during late childhood, peaks during adolescence, and then gradually decreases[11-13]. Together, synaptogenesis and synaptic pruning influence brain plasticity, a feature which allows the brain to adapt in response to experience[14]. Considering synaptic changes are most prominent in infancy, childhood and adolescence, these stages represent periods of enhanced susceptibility to environmental influence, as well as being periods of increased neuropsychiatric risk[15]. Describing the typical cell type-specific gene expression trajectories of the maturing brain will serve as an important reference to assess the effects of genetic perturbations and early adverse experiences on brain maturation. Furthermore, investigating the driving forces behind cell type-specific maturational processes may help in developing targeted therapies for neurological disease[16].

To this end, the Paediatric Cell Atlas (PCA)[17], a branch of the Human Cell Atlas (HCA) efforts to map the cellular diversity of the human body[3], aims to ensure that the benefits of single cell transcriptomics, which will likely include vast improvements in precision medicine, are available to children as well as adults from diverse populations[3,17]. Considering that Africa has the most genetically diverse[18] and youngest population[19] worldwide and that, by 2050, 37% of the world's children will grow up in Africa[20], the inclusion of the African paediatric population in the PCA's efforts has never been more pertinent. A reference paediatric cell atlas for the brain that includes data from African donors will contribute to developing the best treatment for locally prevalent conditions, such as tuberculosis meningitis and HIV infection[21,22]. In addition, studying the differences in gene expression dynamics between adult and paediatric brains may explain why the manifestation of neurological conditions and responses to therapies differ across the lifespan[17].

To contribute to these endeavours, we present a joint paediatric and adult temporal cortex cell atlas, including samples from eight South African donors. To facilitate a direct comparison of our snRNA-seq datasets with the existing adult temporal cortex cell atlas, we annotate our data using the Allen Brain Map middle temporal gyrus (MTG) cell taxonomy[1]. We validate this annotation approach using spatial transcriptomics analysis. In addition, we use *de novo* marker gene analysis with machine learning tools to compare our paediatric and adult datasets to the existing MTG cell taxonomy and highlight markers that define paediatric versus adult cell states. Using differential gene expression analysis, we highlight six cell subtypes, including two layer 2/3 excitatory neuron subtypes, that show differential expression of genes involved in synaptic plasticity and cognition. Overall, we highlight the subtle cell type-specific differences between the paediatric and adult brain and expand the representation of diverse paediatric populations in the HCA.

Results

A joint paediatric and adult temporal cortex cell atlas

We generated snRNA-seq libraries from five paediatric and three adult donor temporal cortex tissue samples using the 10x Genomics Chromium platform. These new libraries were analysed alongside similar published datasets from one paediatric and three adult donors[23], resulting in a total of 23 snRNA-seq datasets (including technical replicates) from 12 individuals (six paediatric and six adult) (Fig. 1a, Extended Data Table 1). The samples were sequenced to a median depth of 19,853 reads per nucleus, with 176,012 nuclei remaining after filtering low quality barcodes (Methods) (Extended data Fig. 1, Extended Data Table 2).

Using data integration and clustering (Methods), we aligned similar cell types across the 23 datasets (Fig. 1a, Extended data Fig. 1h), yielding 40 clusters (Extended data Fig. 1i). To facilitate a direct comparison between our datasets and the current draft human brain cell atlas, we used *Seurat*'s label transfer method[24] to classify each nucleus according to the Allen Brain Map MTG atlas[1] (Fig. 1b; Extended Data Table 3). Of the 75 reference cell types, 21 neuronal subtypes were absent from our datasets (Extended Data Table 3). The majority of the absent neuronal cell types are rare (<2% of all cells; see Extended Data Table 3), and therefore their absence is likely due to the lower proportion of neurons in our datasets compared to the reference atlas, which was made using cell-sorting to enrich for neurons[1].

As an initial validation of the label transfer, we confirmed that the majority of annotated cell types expressed the expected cell type-specific marker genes[1] (Fig. 1c). Additionally, we performed a correlation analysis to compare the transcriptomic similarity of the annotated cell types to the reference MTG cell types[1] (Fig. 1d). The non-neuronal cell types showed high correlation with the corresponding reference cell types, as well as high specificity (Fig. 1d).

Similarly, most neuronal subtypes showed high correlation with the corresponding subtype in the reference datasets, however, there was also correlation with other subtypes within their class. Exceptions were Exc_L4-6_FEZF2_IL26, which expressed the excitatory neuron marker gene *FEZF2* but correlated more strongly with microglia, and Exc_L5-6_THEMIS_FGF10, which expressed the inhibitory marker *LHX6* and correlated more strongly with the inhibitory neuron subtypes (Fig. 1c-d). The overall cell composition of the paediatric and adult samples was very similar. Oligodendrocytes and oligodendrocyte precursor cells (OPCs) were the most common non-neuronal cell types and Exc_L2_LAMP5_LTK was the most common neuronal cell subtype (Fig. 1e; Extended data Figure 2a-c).

Neuronal clusters had a greater number of expressed genes and unique molecular identifiers (UMIs) compared to non-neuronal cells (Extended data Figure 3a). Similar to previous snRNA-seq analyses of the adult brain[1,2], excitatory neurons had a greater number of genes detected per nucleus than inhibitory neurons, with Exc_L2-3_LINC00507_FREM3 and Exc_L4-5_RORB_FOLH1B among the cell types with the highest median gene detection. When comparing the paediatric to adult samples, only two cell subtypes showed significant differences in the number of genes (Exc_L2_LAMP5_LTK and Exc_L3-5_RORB_FILIP1L) and UMIs (Exc_L2_LAMP5_LTK and Exc_L2-3_LINC00507_FREM3) between the age categories, with the paediatric samples having a significantly larger value in each case (Extended data Figure 3b-c). This result points towards higher transcriptional diversity in these neuronal subtypes during childhood.

Spatial mapping of cell types reveals similar tissue cytoarchitecture in adult and paediatric temporal cortex

Next, we used spatial transcriptomics to verify the positions of our annotated cell types within the layered structure of the temporal cortex. We generated 10x Genomics Visium Spatial Gene Expression datasets for frozen tissue sections from adult (31-year-old) and paediatric (15-year-old) temporal cortex samples (two sections each; Extended Data Table 1; Extended data Fig. 4). The four Visium libraries were sequenced to a median depth of 87,178 reads per spot (median of 5,878 UMIs and 2,745 genes per spot) (Extended Data Table 4).

Using *cell2location*[25] (Methods), we calculated cell type abundance estimates for each Visium spot, with our annotated snRNA-seq dataset as a reference. As seen in our snRNA-seq datasets, oligodendrocytes were the most common cell type, while Exc_L2_LAMP5_LTK was the most abundant neuronal cell type (Extended data Fig. 5a). Spatial plots of estimated cell abundance for a selection of cell types revealed that the annotated cell types mapped to their expected cortical layer locations across all tissue sections (Fig. 2a-b; Extended data Fig. 5b-c). The two cell types that did not correlate strongly with the expected MTG reference atlas cell types in our snRNA-seq analysis, Exc_L4-6_FEZF2_IL26 and Exc_L5-6_THEMIS_FGF10, were

most strongly distributed in layers 4-6, providing support for their annotation (Extended data Fig. 5d). The layered pattern of the annotated cell types coincided with the spatial expression of known cortical layer marker genes[1,26,27] in the Visium datasets (Fig. 2c). These layered expression patterns were verified for a subset of layer-specific marker genes using *in situ* hybridisation analysis on frozen temporal cortex tissue sections from the same 31-year-old and 15-year-old donors (Extended data Fig. 6).

To confirm the co-location of cell types within the layered structure of the temporal cortex, non-negative matrix factorization (NMF) was performed using the cell abundance estimates from *cell2location*, resulting in 12 cellular compartments (Fig. 2d). The NMF weights for the identified cellular compartments were visualised across the Visium samples to assess their spatial distribution (Fig. 2e). In both the paediatric and adult datasets, there was clear co-location of the expected neuronal cell types within overlapping compartments across the cortical layers, including layers 1-3 (factor_9), layers 3-5 (factor_3) and layers 4-6 (factor_7). Several excitatory neuron subtypes formed discreet cellular compartments, including Exc_L2-3_LINC00507_FREM (factor_4), Exc_L5-6_THEMIS_CRABP1 (factor_11) and Exc_L5-6_FEZF2_ABO (factor_6). The two astrocyte subtypes were confirmed to have distinct distributions profiles, with Astro_L1-2_FGFR3_GFAP (factor_5) located primarily in layer 1 and the white matter, and Astro_L1-6_FGFR3_SLC14A1 (factor_8) more widely distributed. The remaining non-neuronal cell types were largely associated with factors located in layer 1 and the white matter.

Overall, our spatial transcriptomic analyses provide support for our annotation approach, showing the expected spatial distribution of annotated cell types, and revealing a similar tissue cytoarchitecture in adult and paediatric temporal cortex tissue.

A machine learning approach identifies gene pathways that distinguish paediatric and adult cell states

To establish a standardized and scalable approach for defining cell types, it has been proposed to use the minimum combination of gene markers that can classify a cell type and distinguish it from other cell types[28,29]. Towards achieving this, Aevermann et al. (2021)[28] developed the machine learning tool, *NS-Forest V2.0*, which they applied to the Allen Brain Map MTG dataset. Ideally, theses MTG minimal markers would be conserved in similar datasets to facilitate accurate comparisons across different studies[30]. We found that the MTG cell atlas minimal markers[28] are indeed highly expressed in the expected cell types (Extended data Fig. 7).

Next, we applied the *NS-Forest V2.0*[28] algorithm to our datasets, firstly, to assess if the identified minimal markers overlap with the published MTG minimal markers and, secondly,

to identify combinations of marker genes that distinguish the paediatric and adults states of each cell type (Methods). 151 paediatric and 149 adult minimal marker genes were identified across 53 cell types (Fig. 3; Extended data Table 5). There was little overlap with the MTG atlas, with only 11 paediatric (7.3%) and 4 adult (2.7%) minimal markers showing overlap with existing datasets[1,28] (Fig. 3; Extended data Table 5). On the other hand, there was a greater overlap in minimal markers between the paediatric and adult datasets, with 35 markers (~23%) present in both lists.

Our minimal marker analysis revealed improved markers for several cell types when compared to the reference MTG cell atlas. *DDR2* shows high specificity for Inh_L1-2_PAX6_CDH12 (Fig. 3; Extended data Fig. 8a), while the existing minimal marker for this cell type, *TGFBR2*, is more highly expressed in the microglia and endothelial cells (Extended data Fig. 7; Extended data Fig. 8b). Similarly, *SEMA3E* is very specific to Exc_L5-6_FEZF2_ABO (Fig. 3; Extended data Fig. 8c), while an existing minimal marker for this cell type, *SULF1*, is also expressed at appreciable levels in other neurons (Extended data Fig. 7; Extended data Fig. 8d). Additionally, UMAP analysis of the annotated datasets using our minimal marker gene list for each age group, in comparison to an equivalent number of random genes, resulted in better grouping of the cell subtypes into clusters (Fig. 4 a-b). While the cell clusters are not as clearly separated as the original UMAP plot, (see Fig. 1a), this analysis reveals that our short list of ~150 marker genes captures much of the underlying transcriptomic diversity in our datasets.

To explore the cellular functions of the paediatric and adult minimal marker genes respectively, we used Gene ontology (GO) analysis (Extended Data Table 6), which revealed significant enrichment of GO terms related to neuronal development and cell signalling for both datasets. Interestingly, only the paediatric dataset was enriched for cellular migration terms (Extended Data Table 6). Genes included in these sets were *RELN*, *CXCL14* and *SEMA3A*, which play roles in neuronal migration during brain development[31-33]. On the other hand, only the adult datasets were enriched for extracellular matrix and cell death terms (Extended Data Table 6). These broad analyses of minimal marker gene function indicate that genes involved in neuronal development pathways remain key to neuronal identity, alongside functional signalling molecules, as the brain matures and in adult life. Our results also suggest that genes involved in cellular migration processes may continue to define cell states during childhood and adolescence.

To further assess the difference in cell type-specific markers between our paediatric and adult datasets, we expanded our analysis to include the top 50 genes identified as cell type classification features for each cell type. For most cell types, the majority of these top markers (>20 genes) were shared between our paediatric and adult datasets (Fig. 4c; Extended Data Tables 7-8). The non-neuronal cell types showed the highest number of shared marker genes (≥40). GO analysis of these shared marker genes showed significant enrichment of terms

related to cell type-specific functions, such as "leukocyte proliferation" for microglia, and "myelin sheath" for oligodendrocytes (Extended Data Table 6). These results suggest that paediatric and adult non-neuronal cell states are relatively similar, although it is likely that more diversity in the marker gene profiles could be revealed with subdivision of the cells into further subtypes.

The cell types with the fewest shared markers between our paediatric and adult datasets (≤10) were mostly represented by fewer than 20 nuclei and therefore the lack of marker overlap between our datasets is possibly due to sampling bias (Extended Data Table 6). An exception was Exc_L4-6_FEZF2_IL26, represented by 2,579 nuclei, which had no shared genes, two adult-specific genes (*SNAP25* and *CALM1*) and four paediatric-specific genes (*MEF2A*, *MEF2C*, *DOCK4* and *PLXDC2*) (Fig. 4c; Extended Data Tables 7-8). Interestingly, the paediatric-specific markers for this cell type are implicated in neuronal development processes, including synaptic plasticity[34], dendritic branching[35] and neuronal proliferation[36]. The cell type with the greatest number of paediatric-specific marker genes was Exc_L4-5_RORB_DAPK2 (29 paediatric vs 3 adult markers) (Fig. 4c; Extended Data Tables 7-8). GO analysis of the 29 paediatric-specific marker genes for this cell type revealed enriched terms related to synapse development, while similar analysis of the 19 shared genes together with the three adult-specific genes highlighted terms involved in synaptic function (Extended Data Table 6).

Overall, our expanded marker gene analysis suggests that neuronal cell types show greater dissimilarity between their paediatric and adult states than non-neuronal cells, and reveals several paediatric-specific markers that reflect the less mature state of paediatric neuron subtypes.

Differential gene expression analysis highlights six cell subtypes and enriched expression of genes associated with synaptogenesis and cognition in paediatric samples.

While our minimal marker gene analysis provides insight into the genetic signatures that distinguish cell types from one another over the course of brain maturation, we were also interested in investigating how the general transcriptomic profile of each cell type differs in children when compared to adults. In particular, we sought to identify genes that were upregulated in the paediatric cell populations and thus might be involved in childhood brain development and function. To this end, we conducted cell type-specific differential gene expression analysis with *DESeq2*[37] (Methods).

In total, we detected 166 unique significantly differentially expressed genes (DEGs) across 12 out of the 54 annotated cell types, with some DEGs associated with multiple cell types (Fig. 5a; Extended data Table 9). When the magnitude of fold change of the DEGs was considered,

143 genes across six cell types changed by at least 10%, with a positive fold change indicating higher expression in our paediatric samples (Fig. 5b-c). The six cell types included four excitatory neuron subtypes, one inhibitory subtype and one astrocyte subtype (Fig. 5b-c; Extended data Table 10). For the majority of DEGs, the change in expression was accompanied by a corresponding change in the percentage of nuclei expressing the gene (Extended data Table 10). We assessed the expression patterns of a subset of DEGs in our Visium datasets using *BayesSpace*[38] (Methods), confirming that the genes were expressed at higher levels and in a greater number of spots in the 15-year-old compared to the 31-year-old (Extended data Fig. 9).

The layer 2/3 excitatory neuron subtypes, Exc_L2_LAMP5_LTK and Exc_L2-3_LINC00507_FREM3, shared several upregulated DEGs that are developmentally regulated in the mammalian brain (Fig. 5c-d). *FNBP1L* (*TOCA-1*) regulates neurite outgrowth[39] and is associated with intelligence[40]. In line with our findings, its expression declines over the course of brain maturation in the rat[39]. Similarly, both *KCNG1,* a voltage gated-potassium channel (Kv6.1), and *MYO16* (*MYR8*), an unconventional myosin protein, decrease in expression with age in the mouse[41] and rat[42] brain, respectively. These findings indicate that previously reported expression dynamics for these genes in mammalian models are conserved in the human temporal cortex, with higher expression in children. Importantly, our analysis reveals that these patterns are specific to two layer 2/3 excitatory neuron subtypes.

Two significantly upregulated genes in the layer 2/3 excitatory neurons, *LUZP2* and *RERGL* (Fig. 5c-d), are associated with neuropsychiatric disorders, as well as brain cancers. *LUZP2*, a leucine zipper protein, is associated with Alzheimer's Disease and schizophrenia, as well as cognitive performance in the normal elderly population[43]. It is differentially expressed in several cancers, including low grade gliomas[44]. Similarly, *RERGL* is differentially expressed in schizophrenia[45], while its expression is significantly higher in meningiomas compared to healthy tissue[46]. Our results suggest that these genes likely play a role in the maturation of layer 2/3 excitatory neurons.

The layer 3-5 excitatory neuron subtypes, Exc_L3-5_RORB_ESR1 and Exc_L4-5_RORB_DAPK2, shared 3 upregulated DEGs, which have known roles in the development and reorganisation of neuronal connections (Fig. 5c,e). *XKR4* is a member of a plasma membrane protein family involved in signalling cellular compartments for engulfment, such as during synaptic pruning[47]. These proteins decrease in expression with age in the mouse brain[48]. *TENM1*, is a member of the teneurin transmembrane protein family that regulate cytoskeletal organisation and neurite outgrowth, as well as shaping synaptic connections[49-51]. *AGBL1* (*CCP4*) is a glutamate decarboxylase that mediates deglutamylation of tubulin[52]. This process is essential for the maintenance of neuronal tubulin, and when disrupted, contributes to neurodegeneration[52,53]. Our results point towards a role for these genes specifically in Exc_L3-

5_RORB_ESR1 and Exc_L4-5_RORB_DAPK2 neurons, where they possibly contribute to the re-shaping of neuronal connections during brain maturation.

The majority of the DEGs in the four neuronal subtypes were not shared across the cell types and represent promising candidates for future explorations into molecular mechanisms guiding cell type-specific brain maturation. For example, *FGF13* (*FHF2*) was significantly upregulated in Exc_L3-5_RORB_ESR1 (Fig. 5c). *FGF13* decreases in expression with age in the mouse brain, where it regulates post-natal neurogenesis[54] and axonal formation[55]. Similarly, *PLPPR1* (*PRG1*) was significantly upregulated in Exc_L2-3_LINC00507_FREM3 (Fig. 5c). *PLPPR1* is higher in the postnatal mouse hippocampus than in the adult[56] and is also known to regulate axon growth by modulating cytoskeletal dynamics[57].

In Inh_L2-6_VIP_QPCT, a single lncRNA, *LINC00276*, was upregulated (Fig. 5b). While this non-coding RNA has been shown to be expressed in the brain[58], nothing is known of its function there. Similarly, in Astro_L1-6_FGFR3_SLC14A1, a single non-coding gene, *AC109439.2* (*CTB-1I21*), was significantly upregulated (Fig. 5c, f). This lncRNA has recently been identified as a protective factor in oesophageal cancer[59] and glioma[60]. One of the downregulated genes in this cell type, *ADAM28*, a metalloproteinase, has been found to be upregulated in breast and lung cancers[61]. These results provide new molecular candidates to expand our understanding of molecular mechanisms of astrocyte maturation.

Genes associated with intelligence quotient (IQ) and educational attainment (EA) have recently been shown to be enriched in adult temporal lobe cortical neurons, especially the Exc_L2-3_LINC00507_FREM3 subtype[62]. Since childhood is a key period of cognitive development[63], we explored whether the same genes were found amongst our DEGs. Of the 137 DEGs found in at least one neuronal cell type, 15 (11%) are known to be significantly associated with EA[64] and 4 (3%) with IQ[65]. These included several genes enriched in paediatric samples, such as *MYO16*, *KCNG1* and *LUZP2* (Extended data Table 10).

Overall, our differential expression analysis highlights six cell subtypes that show significant changes in gene expression between children and adults. Several of the genes that are upregulated in in children have known roles in brain development and have been associated with cognitive ability. Our analysis builds on this knowledge by implicating specific paediatric cell subtypes and provides new candidate genes that likely contribute to cell type-specific maturation processes.

Gene pathways involved in synaptic development and functioning are enriched in paediatric cell types

We next used gene set enrichment analysis (GSEA) to conduct a broad analysis of the gene pathways that are differentially regulated across all brain cell types during brain maturation

(Methods). 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 DEGs[66].

In total, 2,003 GOBP terms where enriched in the paediatric samples compared to the adults, while 866 were depleted ($p<0.01$ and $q<0.1$) (Extended data Table 11). When focusing on the 25 most frequently enriched terms, the majority (11 terms) were associated with cellular respiration pathways (Fig. 6; Extended data Table 11). Six of the most commonly enriched terms were linked to synaptic development and functioning. A similar trend was observed when focussing on the cell types from our differential expression analysis, with gene expression, cellular respiration and synapse development pathways dominating the top enriched terms (Fig. 5g).

Focusing on pathways that are depleted in the paediatric brain, four of the top ten depleted terms, including the top term, were associated with neuronal ensheathment (Fig. 6). Interestingly, none of these terms were significantly enriched in oligodendrocytes or OPCs, while they were associated with neuronal sub types, astrocytes and microglia. The remainder of the top ten depleted terms included neuronal morphogenesis, cell adhesion and gene expression pathways.

Overall, our GSEA analysis points towards putative genetic pathways that may drive the differences in synaptic plasticity between the paediatric and adult brain. Pathways related to reorganising and strengthening synapses may be enhanced across multiple cell types during childhood, while those required to limit synaptic growth, such as axonal ensheathment pathways, may need to be suppressed.

Discussion

The brain is the most complex organ in the human body, which continuously changes as we mature and age. Existing studies exploring the transcriptomic changes across the full span of brain maturation have used bulk transcriptomic techniques[6-10], which drown out the subtle molecular events taking place within specific cell types. Here, we unmask these processes, using single cell transcriptomics to compare similar cell types between paediatric and adult datasets.

To facilitate accurate comparisons of cell types across age groups, we used the existing Allen Brain Map MTG cell atlas[1] to annotate our datasets. This demonstrated that the reference atlas, generated from eight adult snRNA-seq datasets, is indeed generalisable[30], and can be used to classify cell types in similar datasets from samples of different ages. This generalisability is essential for healthy human reference atlases to serve as a baseline to

improve our understanding of human development and disease[3]. Our machine-learning marker gene analysis also shows that while the cell type classifications, which are based on the expression of thousands of genes, can be transferred onto new datasets, the minimal markers that define the cell types do vary across datasets. Very few of our *NS-Forest* minimal markers overlap with the existing MTG cell atlas minimal markers[1,28] and some provide better discrimination between cell type than the existing markers. These results highlight that the minimal markers that currently define the MTG cell taxonomy will likely need to be revised as more samples are made available to ensure that the cell type classification is as widely applicable as possible.

Similar to previous analyses of aging in the mouse[66], our integrated analysis of paediatric and adult datasets showed there is little change in cell type composition within the temporal cortex during human brain maturation. However, both our minimal marker analysis and differential gene expression analysis highlight differences in cell states between paediatric and adult cell types. GO analysis of our minimal marker genes revealed a signal for migratory pathways amongst the paediatric minimal markers, including genes that code for chemoattractants, such as *RELN* and *CXCL14*. While these genes are expressed in the mammalian adult brain[26,67,68], our analysis suggests that they play a greater role in defining cell identity in childhood, where these signals possibly function to mediate processes such as dendritic outgrowth and the rearrangement of neuronal synaptic connections.

In recent years, the excitatory pyramidal neurons in the supragranular layers of the MTG have been shown to have high transcriptional diversity[1,69] and to possess unique features including exceptionally large arborisations[70] and electrophysiological properties that impact signal integration and encoding[71-74] in ways that may enhance neuronal computational abilities, contributing to human cognition. Since cognitive ability is a key feature that is established during childhood[68], our differential expression analysis offers an extraordinary opportunity to explore how cell type-specific gene expression dynamics contribute to cognitive development. We highlight six cell subtypes that show significant gene expression changes between childhood and adulthood. Interestingly, two of these cell types were the layer 2/3 excitatory neurons, Exc_L2_LAMP5_LTK and Exc_L2-3_LINC00507_FREM3, that have recently been associated with human cognition[62]. In line with these findings, several of the DEGS shared by these cell types, including *FNBP1L*[40] and *LUZP2*[43], have been implicated in cognitive ability and intelligence. Overall, our data points towards genes and pathways that likely play key roles in cognitive development specifically within these layer 2/3 excitatory neurons.

The relatively low number of genes and cell types implicated in our differential expression analysis in comparison to similar studies in mouse[66] suggests that the difference between the paediatric and adult brain are subtle. However, the inherent high variability in human gene expression data may potentially mask some of the differential gene expression in our limited sample. As the HCA database for the human temporal cortex expands, it will be important to

build on these analyses with more samples for each developmental stage to provide more support for our findings.

We have provided the first single cell gene expression datasets for the brain that includes data from black South African donors, thus increasing the diversity of the HCA database. Our paediatric datasets will form important baseline references for future studies aiming to explore how locally important challenges to child brain health, including infectious diseases and traumatic brain injury, impact on the normal gene expression profiles. Importantly, these investigations will contribute to the development of effective treatments, that are tailored to specific needs of paediatric patients.

Methods

Human samples

Ethical approval was granted for the collection and use of paediatric and adult human brain tissue by the University of Cape Town Human Research Ethics Committee (UCT HREC REF 016/2018; sub-studies 146/2022 and 147/2022). The human brain tissue samples used to generate new 10x Genomics snRNA-seq and Visium datasets were obtained by informed consent for studies during temporal lobe surgical resections to treat epilepsy and/or cancer performed at the Red Cross War Memorial Children's Hospital and Constantiaberg Mediclinic in Cape Town, South Africa. The samples used in this study were of temporal cortex origin and represent radiologically and macroscopically normal neocortex within the pathological context (details in Extended Data Table 1). Upon resection, samples were placed in carbogenated ice-cold artificial cerebral spinal fluid (aCSF) containing in (mM): 110 choline chloride, 26 $NaHCO_3$, 10 D-glucose, 11.6 sodium ascorbate, 7 $MgCl_2$, 3.1 sodium pyruvate, 2.5 KCl, 1.25 NaH_2PO_4, and 0.5 $CaCl_2$ (300 mOsm) and immediately transported to the laboratory (~20 minutes). Tissue blocks containing the full span from pia to white matter were prepared and either flash frozen in liquid nitrogen or embedded in optimal cutting temperature compound (OCT) and stored at -80°C. The OCT-embedded samples were flash frozen in a 10×10 mm^2 cryomold which was either frozen directly in liquid nitrogen or placed in a container of isopentane (Merck) which was in turn placed in liquid nitrogen at the same level as the isopentane. The publicly available snRNA-seq datasets[23], generated from samples obtained during elective surgeries performed at Universitair Ziekenhuis Leuven, Belgium, were downloaded from the Sequence Read Archive database.

Nuclei isolation for snRNA-seq

Nuclei were isolated according to a protocol adapted from Habib et al. (2017)[75] and the 10X Genomics nuclei isolation protocol (CG000124, User Guide Rev E). 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 [Miltenyi 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.

10X Genomics snRNA-seq library preparation

snRNA-seq library preparation was carried out using the 10x Genomics Chromium Next Gen Single Cell 3' Reagent Kit (v3.1) according to manufacturer's protocols (CG000204, User Guide Rev D), targeting 10,000 nuclei per sample. At step 2.2d and 3.5e, the libraries were amplified using 11 cycles and 13 cycles, respectively. Library quality and concentration was assessed using either the TapeStation or Bioanalyser (Agilent) and Qubit (Invitrogen) at the Central Analytical Facility (CAF, University of Stellenbosch). cDNA libraries were sequenced by Novogene (Singapore) on either the Illumina HiSeq or NovaSeq system using the Illumina High Output kits (150 cycles).

snRNA-seq read alignment and gene expression quantification

Fastq files were aligned to the human reference transcriptome (GRCh38) and quantified using the count function from the 10X Genomics Cell Ranger v6.1.1 software (Cell Ranger, RRID SCR_017344) (Code availability: script 1). The inclusion of introns was specified in the count function. An automatic filtering process was performed to remove barcodes corresponding to background noise which have very low UMI counts.

snRNA-seq quality control

The resulting count matrices were processed using a pipeline adapted from the Harvard Chan Bioinformatics Core (https://hbctraining.github.io/scRNA-seq_online/). The filtered gene barcode matrix for each sample was imported into R (V.4.2.0) using the Read10X function from the Seurat (v.2.0)[24]. Nuclei-level filtering was performed to remove poor quality nuclei according to their number of UMIs (nUMIs) detected, number of genes detected (nGene), number of genes detected per UMI (log10GenesPerUMI), and the fraction of mitochondrial read counts to total read counts (mitoRatio) (Code availability: script 2). Nuclei that met the following criteria were retained: nUMI > 500, nGene > 250, log10GenesPerUMI > 0.8 and mitoRatio < 0.2. Gene-level filtering was performed to remove genes that had zero counts in all nuclei, remove genes expressed in fewer than 10 nuclei, and remove mitochondrial genes from the gene by cell counts matrix. Three doublet removal tools namely DoubletFinder[76](Code availability: script 3), DoubletDecon[77] (Code availability: script 4), and Scrublet[77] (Code availability: script 5,6) were used to identify doublets for each dataset individually. The sample-specific parameters of each of the tools were adjusted according to the specified guidelines. To achieve a balance between the false positive and false negative rate of the different doublet detection tools, all doublets identified by DoubletFinder as well as the intersection of the doublets identified by DoubletDecon and Scrublet, were removed[77].

snRNA-seq data normalization, integration and clustering

Principal component analysis was performed to evaluate known sources of within-sample variation between nuclei, namely the mitoRatio and cell cycle phase (Code availability: script 7). The UMI counts of the 3000 most variable features were normalised and scaled on a per sample basis by applying Seurat's SCTransform function with mitoRatio regressed out. A Uniform Manifold Approximation and Projection (UMAP) analysis was performed on the merged object to assess whether integration was necessary. The datasets were subsequently integrated using Seurat's SelectIntegrationFeatures, PrepSCTIntegration, FindIntegrationAnchors, and IntegrateData functions (Code availability: script 7). To cluster the datasets following integration, dimensionality reduction was first performed using UMAP embedding, specifying 40 dimensions (Code availability: script 8). The Seurat FindClusters function was then applied at a resolution of 0.8.

snRNA-seq cluster annotation

Label transfer was performed using Seurat's TransferData function with Allen Brain Map MTG atlas[1] as a reference dataset (Code availability: scripts 9-10). This resulted in each barcode in the query dataset receiving a predicted annotation based on a similarity score to an annotated cell type in the reference. To validate the annotation, the expression of known marker genes was assessed. Additionally, cosine similarity scores were computed to compare the transcriptomic similarity of each of the 54 annotated query cell types to the 75 reference middle temporal gyrus cell types using the SCP package (https://github.com/zhanghao-

njmu/SCP) (Code availability: script 11). This was achieved by computing cosine similarity scores for each pair of query and reference cell types using the expression of the top 2000 shared highly variable features between the query and reference datasets. The log normalised expression counts were used for this purpose (RNA assay, data slot).

NS-Forest machine learning marker analysis of snRNA-seq datasets

The NS-Forest tool (v2.0)[28,29] was used to identify combinations of marker genes uniquely defining each annotated cell type (Code availability: script 12-14) in the paediatric and adult datasets separately. The number of nuclei per sample was randomly down-sampled to that of the sample with the fewest nuclei (n=4865). A random-forest model was used to select a maximum of 50 marker genes per cell type based on them being both highly expressed as well as uniquely expressed within a cell type compared to other cell types (i.e., the top Gini Index ranked features with positive expression values). The number of trees chosen for this model was 35,000, the cluster median expression threshold was set to the default value of zero, the number of genes used to rank permutations of genes by their F-beta-score was 6, and the beta weight of the F score was set to 0.5 allowing the outputs to be directly compared to the Allen Brain Map MTG atlas minimal markers[1]. To assess the relevance of these markers in terms of their capacity to distinguish different cell types in a UMAP analysis, the SCT and integration methods were repeated using either a random set of genes or the NS-Forest markers as anchors[28] (Code availability: script 15).

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

DESeq2[37] was used to identify genes that were differentially expressed with age (Code availability: script 16). 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 Log2 Fold Change = 0). 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 and Hochberg method. Positive log2 Fold Changes represent genes which are upregulated in paediatric samples compared to adult samples (p_{adj} <0.05).

Pathway enrichment analysis of snRNA-seq datasets

GO analysis of NS-Forest marker genes was performed on the gProfiler web server[78] using default settings (p_{adj} <0.05) with "highlight diver terms in GO" selected.

Neuronal DEGs identified by DESeq2 (see Extended data Table 10) that were associated with EA and IQ were determined by comparing the list of neuronal DEGS to the EA and IQ gene lists used by Driessens et al. (2023)[62], which were subsets of the lists from Lee et al. (2018)[64] and Savage et al. (2018)[65] respectively.

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 17). 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-value)*log_2FoldChange 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 abs(log_2FC)>0.1], the top 5 positively associated terms were plotted.

snRNA-seq data plots

Plots were produced with Seurat[24], ggplot2[79], ShinyCell[80] and Microsoft Excel.

10x Genomics Visium library preparation

Frozen OCT embedded brain tissue samples were scored using a pre-chilled razor blade to fit in the Spatial Gene Expression slide capture areas. 10 µm-thick sections were cut using a cryostat (Leica CM1860/CM1950) and collected onto the Spatial Gene Expression slide capture areas. Two replicate sections of the 15-year-old (10 µm apart) and two replicate sections of 31-year-old (40 µm apart) were collected. The spatial Gene Expression slides with tissue sections were stored in a sealed container at -80°C. Captured sections were Haematoxylin and Eosin (H&E) stained according to the 10x Genomics Demonstrated Protocol Guide (CG000160, Rev B). Brightfield images of the stained sections were captured using an EVOS M5000 microscope (Thermo Fisher Scientific) at 20x magnification without coverslipping. Overlapping images of the sections including the fiducial frame were stitched together using Image Composite Editor-2.0.3 (Microsoft). Visium libraries were prepared

from the stained tissue sections following the Visium Spatial Gene Expression Reagents Kit User Guide (CG000239, Rev D). At Step 1.1 the tissue was permeabilised for 12 minutes as determined using the Visium Spatial Gene Expression Tissue Optimisation User Guide (CG000238, Rev D). At Step 3.2, cDNA was amplified using 20 cycles. Library quality and concentration was assessed using TapeStation (Agilent) and Qubit (Invitrogen) at the Central Analytical Facility (CAF, University of Stellenbosch). Libraries were sequenced by Novogene (Singapore) on the Illumina NovaSeq system using the Illumina High Output kits (150 cycles).

Visium read alignment and gene expression quantification

The H&E images were processed using the 10X Genomics Loupe Browser V4.0 Visium Manual Alignment Wizard. 10X Genomics Space Ranger *count* (10X Space Ranger V1.3.0) was used to perform alignment of FASTQ files to the human reference transcriptome (GRCh38), tissue detection, fiducial detection and barcode/UMI counting.

cell2location analysis of Visium datasets

The average number of nuclei per Visium spot was determined using Vistoseg[81] (Code availability: script 18). Cell2location (version 0.7a0)[25] was used to spatially map the brain cell types by integrating the Visium data count matrices (Space Ranger output) with the annotated snRNAseq datasets (Code availability: script 19). To avoid mapping artifacts, mitochondrial genes were removed from the Visium datasets prior to spatial mapping. Reference signatures of the 54 annotated cell populations were derived using a negative binomial regression model using the default values. Unnormalized and untransformed snRNA-seq mRNA counts were used as input in the regression model for estimating the reference signatures (Code availability: script 20). The snRNA-seq mRNA counts were filtered to 13,870 genes and 176,012 cells. The cell2location model for estimating the spatial abundance of cell populations was filtered to 13,858 genes and 14,324 cells that were shared in both the snRNA-seq and Visium data. The following cell2location parameters were used: training iterations = 30,000 cell per location, N^ = 10 (estimated using Vistoseg segmentation results), Normalization (ys) alpha prior = 20 (Code availability: script 21). To visualise the cell abundance in spatial coordinates 5 % quantile of the posterior distribution was used, which represents the value of cell abundance that the model has high confidence in (Code availability: script 22). Cell2location's Non-negative Matrix Factorization (NMF) was used to identify cellular compartments and cell types that co-locate from the cell type abundance estimates, using n_fact=12 (Code availability: script 23)

BayesSpace analysis of Visium datasets

The raw gene expression counts from Space Ranger were normalized, log transformed and principal component analysis was performed on the top 2000 highly variable genes. To obtain high-resolution gene expression, the principal component values were mapped back to their original log-transformed gene expression space (spot level) using the default BayeSpace[38] regression (Code availability: script 24). To do this the principal components from the original

data were used as predictors in training the model for each gene, in which the results were the measured gene expression at the spot level. The trained model was then used to predict the gene expression at sub spot level using high resolution PCs. The high-resolution model was trained using default values except for the following parameters: 7 PCs, Number of clusters = 8, nrep = 100,000, burn-in = 10,000.

***In situ* Hybridisation Chain Reaction (HCR) on frozen human tissue sections**

10 µm thick frozen sections were collected on Histobond+ slides (Marienfeld) and stored at -20°C. The *In situ* HCR protocol was carried out on tissue sections as detailed in Choi et al. (2016)[82] using reagents, probes and hairpins purchased from Molecular Instruments. Probes were ordered for the following genes: *RELN* (NM_005045.4), *FABP7* (CR457057.1), *AQP4* (NM_001650.5), *RORB* (NM_006914.4), *CLSTN2* (NM_022131.3) and *TSHZ2* (NM_173485.6). When necessary to quench lipofuscin autofluorescence, sections were rinsed after HCR in 1x PBS and treated with 200 µl TrueBlack (Biotium) for 30 sec. Slides were rinsed in PBS, stained with Hoescht (Thermofisher) and mounted using SlowFade Gold Antifade Reagent (Invitrogen). Sections were imaged using the LSM 880 Airyscan confocal microscope (Carl Zeiss, ZEN SP 2 software) using the 40X or 60X objective.

References

1 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68 (2019).

2 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119, doi:10.1038/s41586-021-03465-8 (2021).

3 Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, doi:10.7554/eLife.27041 (2017).

4 Network, B. I. C. C. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86-102, doi:10.1038/s41586-021-03950-0 (2021).

5 Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290, doi:10.1073/pnas.1507125112 (2015).

6 Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, doi:10.1126/science.aat7615 (2018).

7 Werling, D. M. *et al.* Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. *Cell Rep* **31**, 107489, doi:10.1016/j.celrep.2020.03.053 (2020).

8 Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489, doi:10.1038/nature10523 (2011).

9 Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519-523, doi:10.1038/nature10524 (2011).

10 Donertas, H. M. *et al.* Gene expression reversal toward pre-adult levels in the aging human brain and age-related loss of cellular identity. *Sci Rep* **7**, 5894, doi:10.1038/s41598-017-05927-4 (2017).

11 Bourgeois, J. P. & Rakic, P. Changes of synaptic density in the primary visual cortex of the macaque monkey from fetal to adult stage. *J. Neurosci.* **13**, 2801-2820, doi:10.1523/JNEUROSCI.13-07-02801.1993 (1993).

12 Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* **387**, 167-178, doi:10.1002/(sici)1096-9861(19971020)387:2<167::aid-cne1>3.0.co;2-z (1997).

13 Petanjek, Z. *et al.* Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc Natl Acad Sci U S A* **108**, 13281-13286, doi:10.1073/pnas.1105108108 (2011).

14 Fu, M. & Zuo, Y. Experience-dependent structural plasticity in the cortex. *Trends Neurosci.* **34**, 177-187, doi:10.1016/j.tins.2011.02.001 (2011).

15 Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience* **9**, 947-957, doi:10.1038/nrn2513 (2008).

16 Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117-1125, doi:10.1038/s41593-018-0197-y (2018).

17 Taylor, D. M. *et al.* The pediatric cell atlas: defining the growth phase of human development at single-cell resolution. *Dev. Cell* **49**, 10-29 (2019).

18 Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044, doi:10.1126/science.1172257 (2009).

19 PRB. *World Population Datasheet 2022*, <https://www.prb.org/wp-content/uploads/2022/09/2022-World-Population-Data-Sheet-Booklet.pdf> (2022).

20 O'Malley, J., Wardlaw, T., You, D., Hug, L. & Anthony, D. Africa's child demographics and the world's future. *Lancet* **384**, 730-732, doi:10.1016/S0140-6736(14)61331-3 (2014).

21 Schutte, C. M. Analysis of HIV-related mortality data in a tertiary South African neurology unit, 2006-2012. *Southern African Journal of HIV Medicine* **14**, 121-124 (2013).

22 Rohlwink, U. K. *et al.* Clinical characteristics and neurodevelopmental outcomes of children with tuberculous meningitis and hydrocephalus. *Dev Med Child Neurol* **58**, 461-468, doi:10.1111/dmcn.13054 (2016).

23 Thrupp, N. *et al.* Single-nucleus RNA-Seq is not suitable for detection of microglial activation genes in humans. *Cell reports* **32**, 108189 (2020).

24 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821, doi:https://doi.org/10.1016/j.cell.2019.05.031 (2019).

25 Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661-671, doi:10.1038/s41587-021-01139-4 (2022).

26 Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425-436 (2021).

27 Hintsch, G. *et al.* The calsyntenins—a family of postsynaptic membrane proteins with distinct neuronal expression patterns. *Mol. Cell. Neurosci.* **21**, 393-409 (2002).

28 Aevermann, B. *et al.* A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. *Genome Res.* **31**, 1767-1780 (2021).

29 Aevermann, B. D. *et al.* Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* **27**, R40-R47, doi:10.1093/hmg/ddy100 (2018).

30 Tan, S. Z. K. *et al.* Brain Data Standards - A method for building data-driven cell-type ontologies. *Sci Data* **10**, 50, doi:10.1038/s41597-022-01886-2 (2023).

31 Tissir, F. & Goffinet, A. M. Reelin and brain development. *Nat. Rev. Neurosci.* **4**, 496-505, doi:10.1038/nrn1113 (2003).

32 Park, C. R. *et al.* Spatiotemporal expression and functional implication of CXCL14 in the developing mice cerebellum. *Mol Cells* **34**, 289-293, doi:10.1007/s10059-012-0116-0 (2012).

33 Tamamaki, N., Fujimori, K., Nojyo, Y., Kaneko, T. & Takauji, R. Evidence that Sema3A and Sema3F regulate the migration of GABAergic neurons in the developing neocortex. *J. Comp. Neurol.* **455**, 238-248 (2003).

34 Akhtar, M. W. *et al.* In vivo analysis of MEF2 transcription factors in synapse regulation and neuronal survival. *PLoS One* **7**, e34863, doi:10.1371/journal.pone.0034863 (2012).

35 Ueda, S., Fujimoto, S., Hiramoto, K., Negishi, M. & Katoh, H. Dock4 regulates dendritic development in hippocampal neurons. *J. Neurosci. Res.* **86**, 3052-3061, doi:10.1002/jnr.21763 (2008).

36 Miller-Delaney, S. F., Lieberam, I., Murphy, P. & Mitchell, K. J. Plxdc2 is a mitogen for neural progenitors. *PloS one* **6**, e14565 (2011).

37 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).

38 Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375-1384 (2021).

39 Kakimoto, T., Katoh, H. & Negishi, M. Regulation of neuronal morphology by Toca-1, an F-BAR/EFC protein that induces plasma membrane invagination. *J. Biol. Chem.* **281**, 29042-29053 (2006).

40 Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Mol. Psychiatry* **19**, 253-258 (2014).

41 Khatir, I. *et al.* Decoupling of mRNA and Protein Expression in Aging Brains Reveals the Age-Dependent Adaptation of Specific Gene Subsets. *Cells* **12**, doi:10.3390/cells12040615 (2023).

42 Patel, K. G., Liu, C., Cameron, P. L. & Cameron, R. S. Myr 8, a novel unconventional myosin expressed during brain development associates with the protein phosphatase catalytic subunits 1alpha and 1gamma1. *J. Neurosci.* **21**, 7954-7968, doi:10.1523/JNEUROSCI.21-20-07954.2001 (2001).

43 Stepanov, V. *et al.* Analysis of Association of Genetic Markers in the LUZP2 and FBXO40 Genes with the Normal Variability in Cognitive Performance in the Elderly. *Int J Alzheimers Dis* **2018**, 2686045, doi:10.1155/2018/2686045 (2018).

44 Feng, D. *et al.* A pan-cancer analysis of the oncogenic role of leucine zipper protein 2 in human cancer. *Exp Hematol Oncol* **11**, 55, doi:10.1186/s40164-022-00313-x (2022).

45 Perez-Santiago, J. *et al.* A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia. *J Psychiatr Res* **46**, 1464-1474, doi:10.1016/j.jpsychires.2012.08.005 (2012).

46 Chen, J. *et al.* Identification of the Key Immune Cells and Genes for the Diagnostics and Therapeutics of Meningioma. *World Neurosurg* **176**, e501-e514, doi:10.1016/j.wneu.2023.05.090 (2023).

47 Maruoka, M. & Suzuki, J. Regulation of phospholipid dynamics in brain. *Neurosci Res* **167**, 30-37, doi:10.1016/j.neures.2021.01.003 (2021).

48 Neniskyte, U. *et al.* Phospholipid scramblase Xkr8 is required for developmental axon pruning via phosphatidylserine exposure. *The EMBO Journal*, e111790 (2023).

49 Zhang, X., Lin, P.-Y., Liakath-Ali, K. & Südhof, T. C. Teneurins assemble into presynaptic nanoclusters that promote synapse formation via postsynaptic non-teneurin ligands. *Nature Communications* **13**, 2297 (2022).

50 Cheung, A. *et al.* Teneurin paralogues are able to localise synaptic sites driven by the intracellular domain and have the potential to form cis-heterodimers. *Front Neurosci* **16**, 915149, doi:10.3389/fnins.2022.915149 (2022).

51 Beckmann, J., Schubert, R., Chiquet-Ehrismann, R. & Müller, D. J. Deciphering teneurin domains that facilitate cellular recognition, cell–cell adhesion, and neurite outgrowth using atomic force microscopy-based single-cell force spectroscopy. *Nano Lett.* **13**, 2937-2946 (2013).

52 Rogowski, K. *et al.* A family of protein-deglutamylating enzymes associated with neurodegeneration. *Cell* **143**, 564-578, doi:10.1016/j.cell.2010.10.014 (2010).

53 Bodakuntla, S., Janke, C. & Magiera, M. M. Tubulin polyglutamylation, a regulator of microtubule functions, can cause neurodegeneration. *Neurosci. Lett.* **746**, 135656, doi:10.1016/j.neulet.2021.135656 (2021).

54 Yang, Q. Q. *et al.* Nuclear isoform of FGF13 regulates post-natal neurogenesis in the hippocampus through an epigenomic mechanism. *Cell Rep* **35**, 109127, doi:10.1016/j.celrep.2021.109127 (2021).

55 Wu, Q.-F. *et al.* Fibroblast growth factor 13 is a microtubule-stabilizing protein regulating neuronal polarization and migration. *Cell* **149**, 1549-1564 (2012).

56 Bräuer, A. U. *et al.* A new phospholipid phosphatase, PRG-1, is involved in axon growth and regenerative sprouting. *Nat. Neurosci.* **6**, 572-578 (2003).

57 Agbaegbu Iweka, C., Hussein, R. K., Yu, P., Katagiri, Y. & Geller, H. M. The lipid phosphatase-like protein PLPPR1 associates with RhoGDI1 to modulate RhoA activation in response to axon growth inhibitory molecules. *J. Neurochem.* **157**, 494-507, doi:10.1111/jnc.15271 (2021).

58 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).

59 Zheng, Z.-J. *et al.* Construction of the Six-lncRNA prognosis signature as a novel biomarker in esophageal squamous cell carcinoma. *Frontiers in Genetics* **13**, 839589 (2022).

60 Xing, Z. *et al.* Clinical Significance and Immune Landscape of a Pyroptosis-Derived LncRNA Signature for Glioblastoma. *Front Cell Dev Biol* **10**, 805291, doi:10.3389/fcell.2022.805291 (2022).

61 Mochizuki, S. *et al.* Effect of ADAM28 on carcinoma cell metastasis by cleavage of von Willebrand factor. *J. Natl. Cancer Inst.* **104**, 906-922, doi:10.1093/jnci/djs232 (2012).

62 Driessens, S. L. *et al.* Genes associated with cognitive ability and HAR show overlapping expression patterns in human cortical neuron types. *Nature communications* **14**, 4188 (2023).

63 Gauvain, M. *Cognitive development in infancy and childhood*. (Cambridge University Press, 2022).

64 Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112-1121, doi:10.1038/s41588-018-0147-3 (2018).

65 Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912-919, doi:10.1038/s41588-018-0152-6 (2018).

66 Ximerakis, M. *et al.* Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* **22**, 1696-1708, doi:10.1038/s41593-019-0491-3 (2019).

67 Wu, J., Zhao, Z., Shi, Y. & He, M. Cortical VIP(+) Interneurons in the Upper and Deeper Layers Are Transcriptionally Distinct. *J. Mol. Neurosci.* **72**, 1779-1795, doi:10.1007/s12031-022-02040-8 (2022).

68 Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78, doi:10.1038/s41586-018-0654-5 (2018).

69 Berg, J. *et al.* Human neocortical expansion involves glutamatergic neuron diversification. *Nature* **598**, 151-158, doi:10.1038/s41586-021-03813-8 (2021).

70 Mohan, H. *et al.* Dendritic and Axonal Architecture of Individual Pyramidal Neurons across Layers of Adult Human Neocortex. *Cereb. Cortex* **25**, 4839-4853, doi:10.1093/cercor/bhv188 (2015).

71 Gidon, A. *et al.* Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* **367**, 83-87, doi:10.1126/science.aax6239 (2020).

72 Eyal, G. *et al.* Human Cortical Pyramidal Neurons: From Spines to Spikes via Models. *Front Cell Neurosci* **12**, 181, doi:10.3389/fncel.2018.00181 (2018).

73 Testa-Silva, G. *et al.* High bandwidth synaptic communication and frequency tracking in human neocortex. *PLoS Biol.* **12**, e1002007, doi:10.1371/journal.pbio.1002007 (2014).

74 Beaulieu-Laroche, L. *et al.* Enhanced Dendritic Compartmentalization in Human Cortical Neurons. *Cell* **175**, 643-651 e614, doi:10.1016/j.cell.2018.08.045 (2018).

75 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955-958, doi:10.1038/nmeth.4407 (2017).

76 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).

77 DePasquale, E. A. K. *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep* **29**, 1718-1727 e1718, doi:10.1016/j.celrep.2019.09.082 (2019).

78 Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191-W198, doi:10.1093/nar/gkz369 (2019).

79 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer, 2016).

80 Ouyang, J. F., Kamaraj, U. S., Cao, E. Y. & Rackham, O. J. L. ShinyCell: simple and sharable visualization of single-cell gene expression data. *Bioinformatics* **37**, 3374-3376, doi:10.1093/bioinformatics/btab209 (2021).

81 Tippani, M. *et al.* Vistoseg: a matlab pipeline to process, analyze and visualize high resolution histology images for visium spatial transcriptomics data. *bioRxiv* (2021).

82 Choi, H. M. *et al.* Mapping a multiplexed zoo of mRNA expression. *Development* **143**, 3632-3637, doi:10.1242/dev.140137 (2016).

Data Availability

All scripts used to analyse the data are indicated in the methods section and are available in the supplementary material. A description of the raw and analysed data files will be made available on the University of Cape Town's ZivaHub data sharing platform on publication. As the data is from living donors, access to the data will be mediated through contact with the corresponding author. A ShinyApp will be made publicly available on publication for exploration of the annotated snRNA-seq data.

Acknowledgments

This research was support by a Royal Society/Global Challenges Research Fund/African Academy of Sciences Future Leaders African Independent Researcher Fellowship to DH (FLR\R1\191008), a National Research Foundation (NRF) Research Development Grant for Y-rated Researchers Award to DH (CSRP210415595025), a University of Cape Town (UCT) Building Research Active Academic Staff Grant award to DH and a National Institutes of Health (NIH) R21 Exploratory/Development Grant to DH, JVR, CGD and MJ (TW011225). CS was supported by a Harry Crossley Research Scholarship, an Oppenheimer Memorial Trust scholarship, an NRF scholarship and a UCT Vice Chancellor's Research Scholarship. SF was supported by a DAAD-NRF joint In-country Scholarship and a UCT Vice Chancellor's Research Scholarship. MBV was supported by an EMBO long-term fellowship (ALTF 415-2018) and a Claude Leon Foundation research fellowship. AF was supported by the NRF SARChI Chair of Clinical Neurosciences. Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za. This publication is part of the Human Cell Atlas: www.humancellatlas.org/publications/

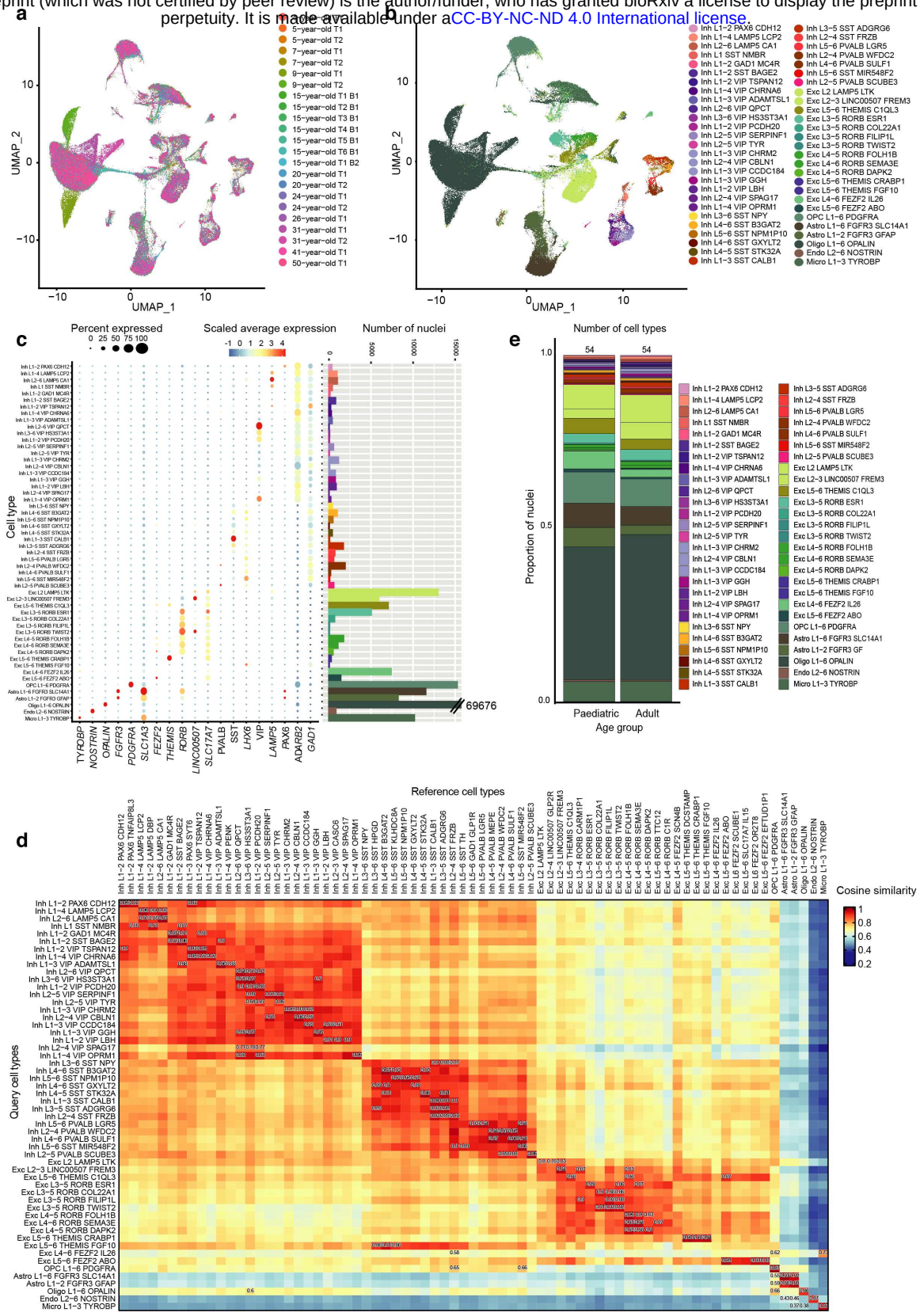


Fig. 1: Annotation of nuclei by label transfer identifies 54 cortical subtypes across the 23 datasets. a, Data integration shows alignment of nuclei across the technical (T) and biological (B) replicates from donors ranging in age from 4 to 50 years. **b**, Annotated UMAP plot for the 23 merged datasets identifies 34 inhibitory, 14 excitatory, and 6 non-neuronal populations using the Allen Brain Map MTG dataset as a reference. Each cell type is annotated with 1) a major cell class (e.g. Exc for excitatory neuron), 2) the cortical layer the cell is associated with (e.g. L2 for layer 2), 3) a subclass marker gene and 4) a cluster-specific marker gene. **c**, Validation of the high-resolution cell type annotations shows a high degree of correspondence in the expression of known cell type-specific markers (x axis) with their expected cell type (y axis) (left). Number of nuclei per cell type (right). **d**, Correlation plot showing the cosine similarity scores assessing similarity between the annotated cell types in our dataset (y axis) and the MTG reference dataset (x axis) based on the log normalized expression counts of the top 2000 shared highly variable features between query and reference datasets. **e**, Stacked barplot showing the proportion of nuclei per cell type (y axis) for each age category (x axis) out of the total number of nuclei for each group. The colour scheme for the cell types is in accordance with the MTG cell taxonomy.

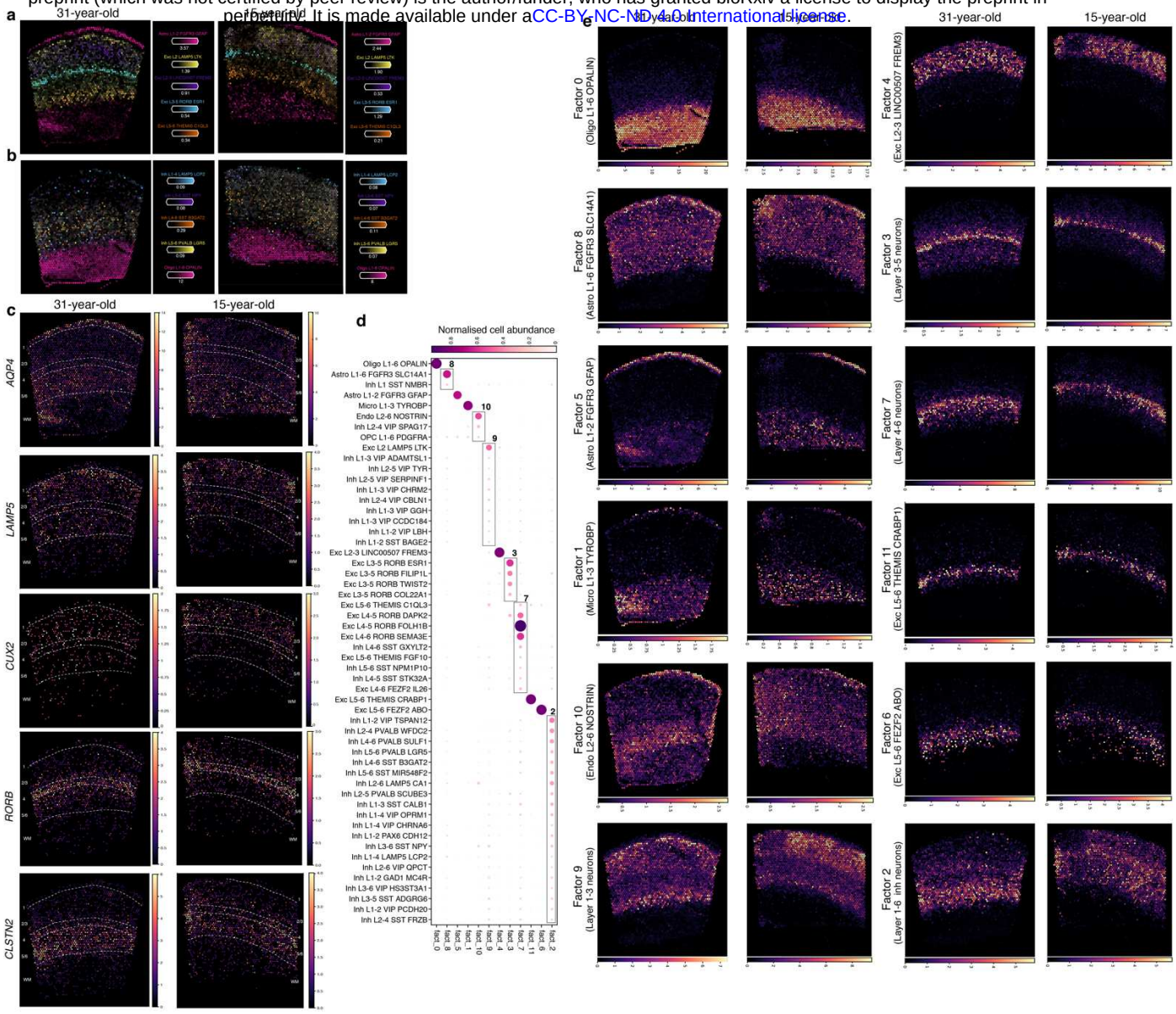


Fig. 2: Fig. 2: Visium spatial transcriptomics in the adult and paediatric temporal cortex validates snRNA-seq annotation. a,**b**, Estimated cell type abundances (colour intensity) in the 31-year-old and 15-year-old temporal cortex tissue sections for a selection of cell types including non-neuronal cell types (**a**,**b**), excitatory neurons (**a**) and inhibitory neurons (**b**). **c**, Visium gene expression profiles (colour intensity) for a selection of known cortical layer marker genes in the 31-year-old and 15-year-old temporal cortex tissue sections including *AQP4* (layer 1), *LAMP5* (layer 2), *CUX2* (layer 2-3), RORB (layer 4) and *CLSTN2* (layer 5-6). Dashed white lines and numbers indicate estimated cortical layer boundaries. **d**,**e**, Identification of co-locating cell types using NMF. The dot plot (**d**) shows the NMF weights of the cell types (rows) across each of the NMF factors (columns), which correspond to tissue compartments. Block boxes indicate cell types that co-locate within the indicated tissue compartments. Spatial plots show (**e**) show the NMF weights for each NMF factor/tissue compartment across the 31-year-old and 15-year-old temporal cortex tissue sections. Panels are displayed in the same order as the dotplot in (**d**), with the dominant cell types for each factor indicated in brackets. WM: white matter. See also Extended Data Figs 4-6.

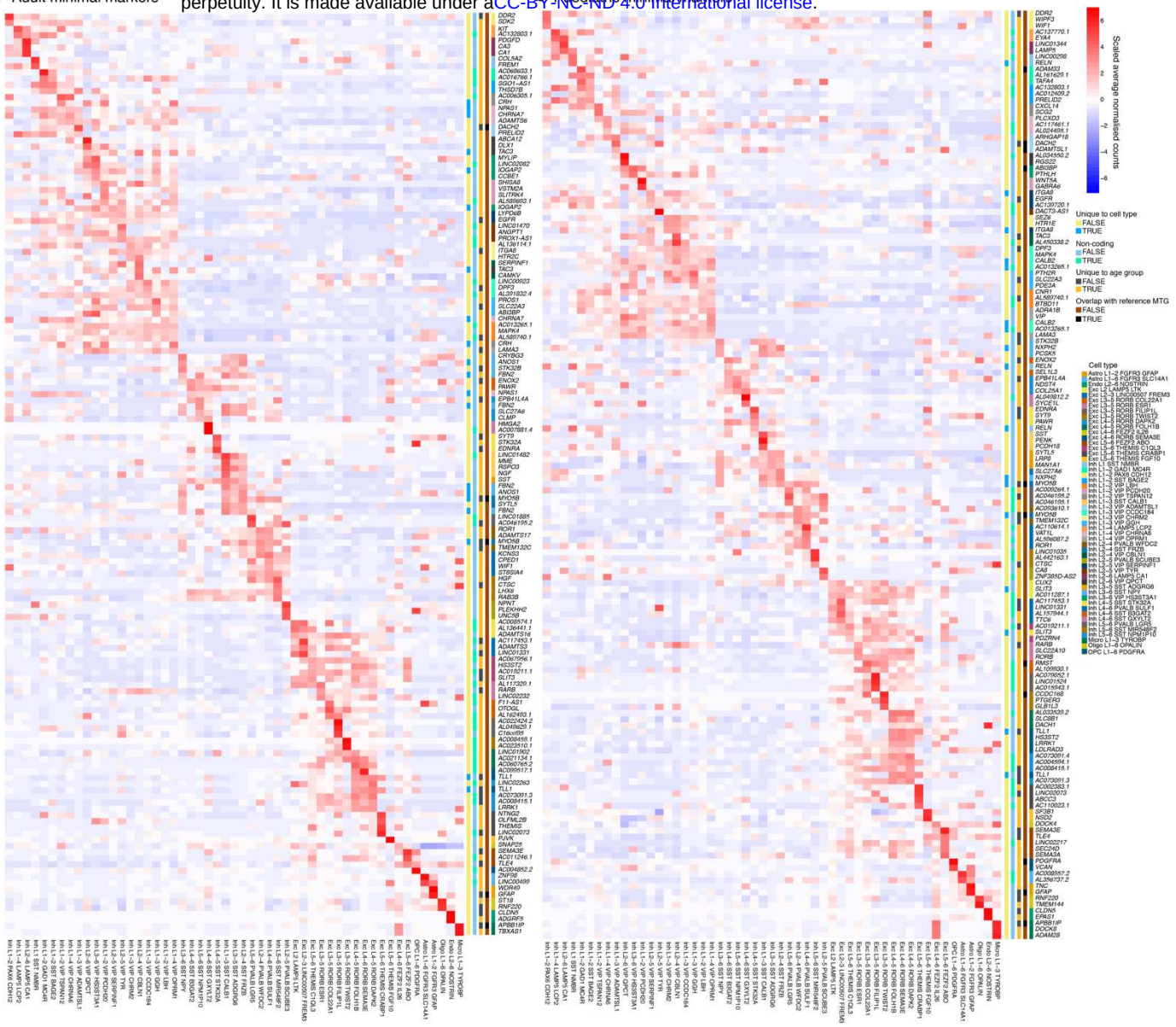


Fig. 3: NS-Forest identifies minimal marker genes distinguishing the cell types in the paediatric and adult temporal cortex snRNA-seq datasets. a,**b**, Heatmap showing the scaled average normalised expression counts of the NS-Forest minimal marker genes (y-axis) identified for 53 cortical cell types (x-axis) across the six adult (**a**) and six paediatric (**b**) datasets. As input into NS-Forest, the nuclei of each sample were randomly down-sampled to the size of the sample with the fewest nuclei. Heatmaps show gene expression values for the down-sampled datasets. Inh_L2−4_VIP_SPAG17 was excluded from the down-sampled datasets due to a low number of nuclei. The minimal marker genes are annotated (colour codes on the y-axes) according to whether they are unique to a given cell type, whether they are coding/non-coding genes, whether they are unique to the indicated age group, whether they overlap with existing MTG minimal marker gene sets for the same cell type, and according to the cell type they define.

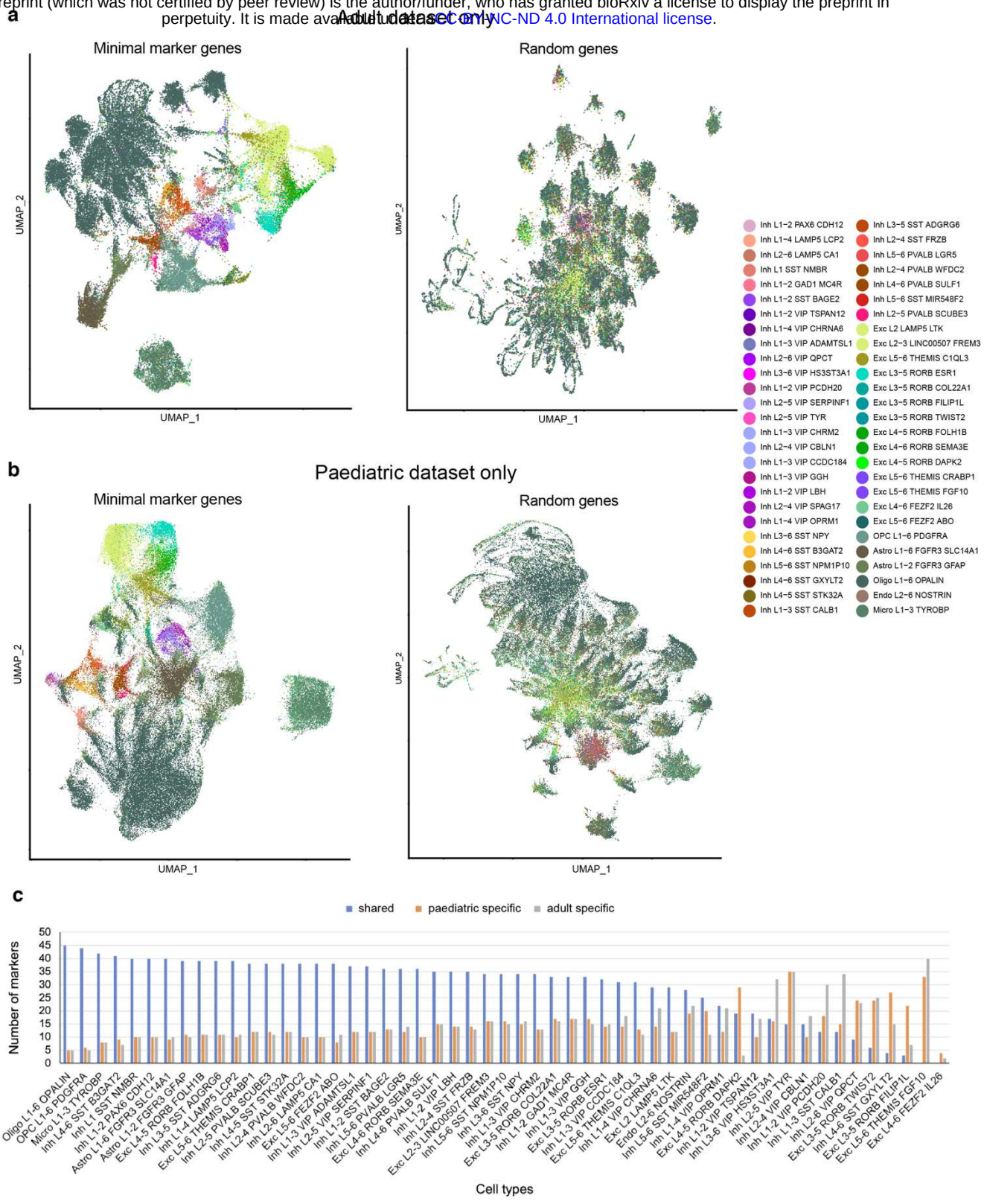


Fig. 4: Validation of NS-Forest minimal markers and assessment of the top 50 NS-forest markers. a,**b**, Annotated UMAP plots following data integration using either the minimal marker genes (left) or the equivalent number of a random set of genes (right) as anchors for the adult (**a**) and paediatric (**b**) datasets. The colour scheme for the cell types is in accordance with the MTG cell taxonomy. **c**, Overlap of the top 50 paediatric and top 50 adult NS-Forest markers per cell type. The bar plot shows the number of shared markers between paediatric and adult datasets (blue), the number of markers unique to the paediatric datasets (orange), and the number of markers unique to the adult datasets (grey) for each cell type.

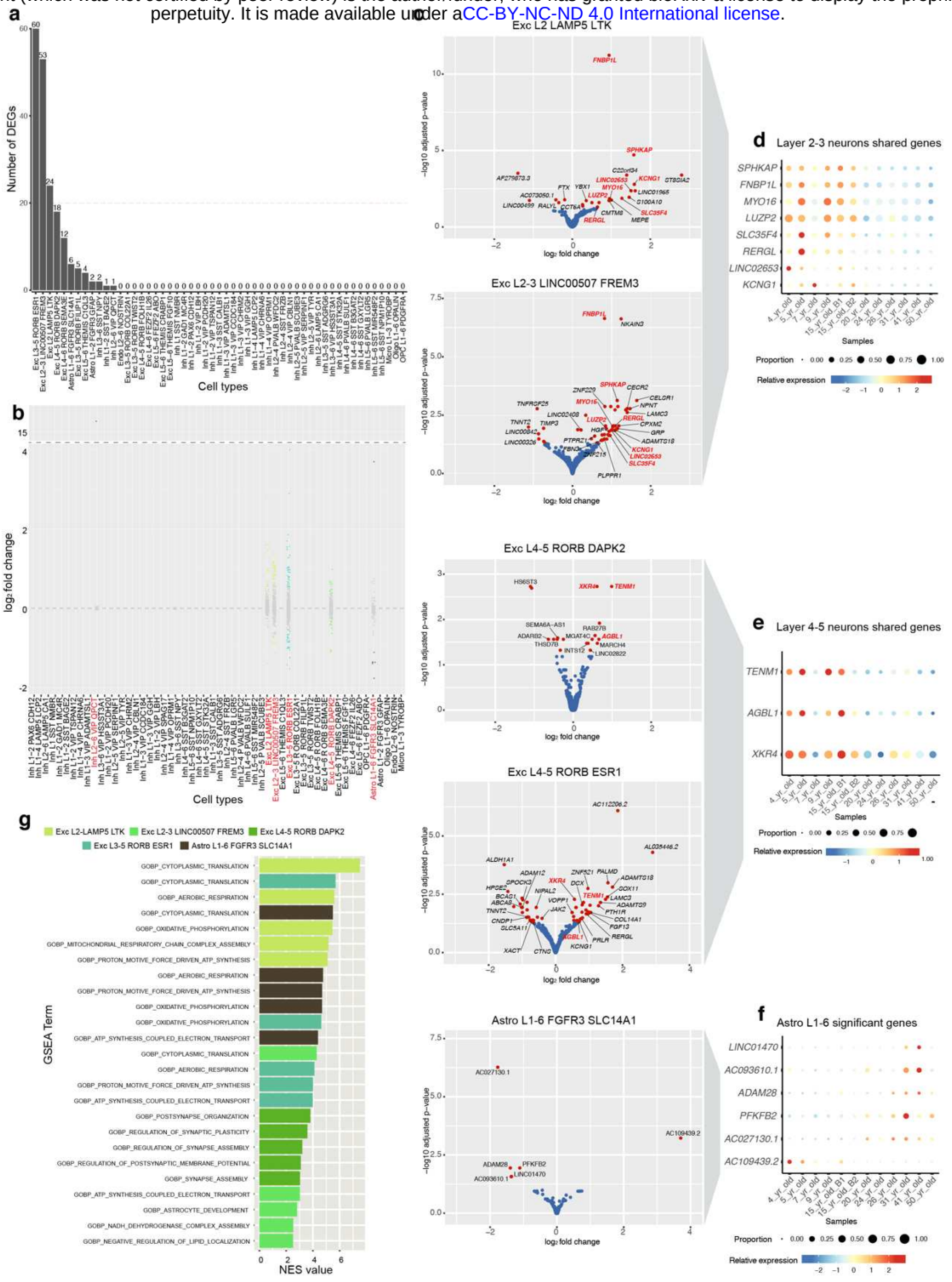


Fig. 5: Differential expression analysis reveals genes guiding temporal cortex maturation. a, The number of differentially expressed genes (DEGs) identified per cell type. **b**, Strip chart showing the log_2FoldChange for all DESeq2-tested genes (dots) across 54 cell types (x axis). Genes that were significantly upregulated (110) or downregulated (52) in paediatric samples [padj<0.05 and abs(log_2FoldChange)>10%)] are coloured according to the MTG cell taxonomy. Non-significant genes [padj>0.05 or abs(log_2FoldChange)<10%] are coloured in grey. Cell types with significant DEGS are coloured red. **c**, Volcano plots showing log_2FoldChange (x axis) and -log_{10}padj values (y axis) for all DESeq2-tested genes in Exc_L2_LAMP5_LTK, Exc L2-3_LINC00507_FREM3, Exc_L4-5_RORB_DAPK2, Exc_L3-5_RORB_ESR1 and Astro_L1-6_FGFR3_SLC14A1. Genes that were significantly upregulated or downregulated in paediatric samples (padj<0.05 & abs(log_2FoldChange)>10%) are coloured in red and selected genes are labelled. Non-significant genes (padj>0.05 or abs(log_2FoldChange)<10%) are coloured in blue. Red labels indicate DEGs shared between similar cell types. **d**,**e**, Dot plots showing the scaled average normalised expression across samples for DEGS shared between either Exc_L2_LAMP5_LTK and Exc L2-3_LINC00507_FREM3 (**d**) or Exc_L4-5_RORB_DAPK2 and Exc_L3-5_RORB_ESR1 (**e**). **f**, Dot plot showing the scaled average normalised expression across samples of DEGs in Astro_L1-6_FGFR3_SLC14A1**. g**, Enrichment plot showing the top 5 GSEA terms (y axis) that were enriched in the paediatric datasets and their associated NES values (x axis) for the five indicated cell types.

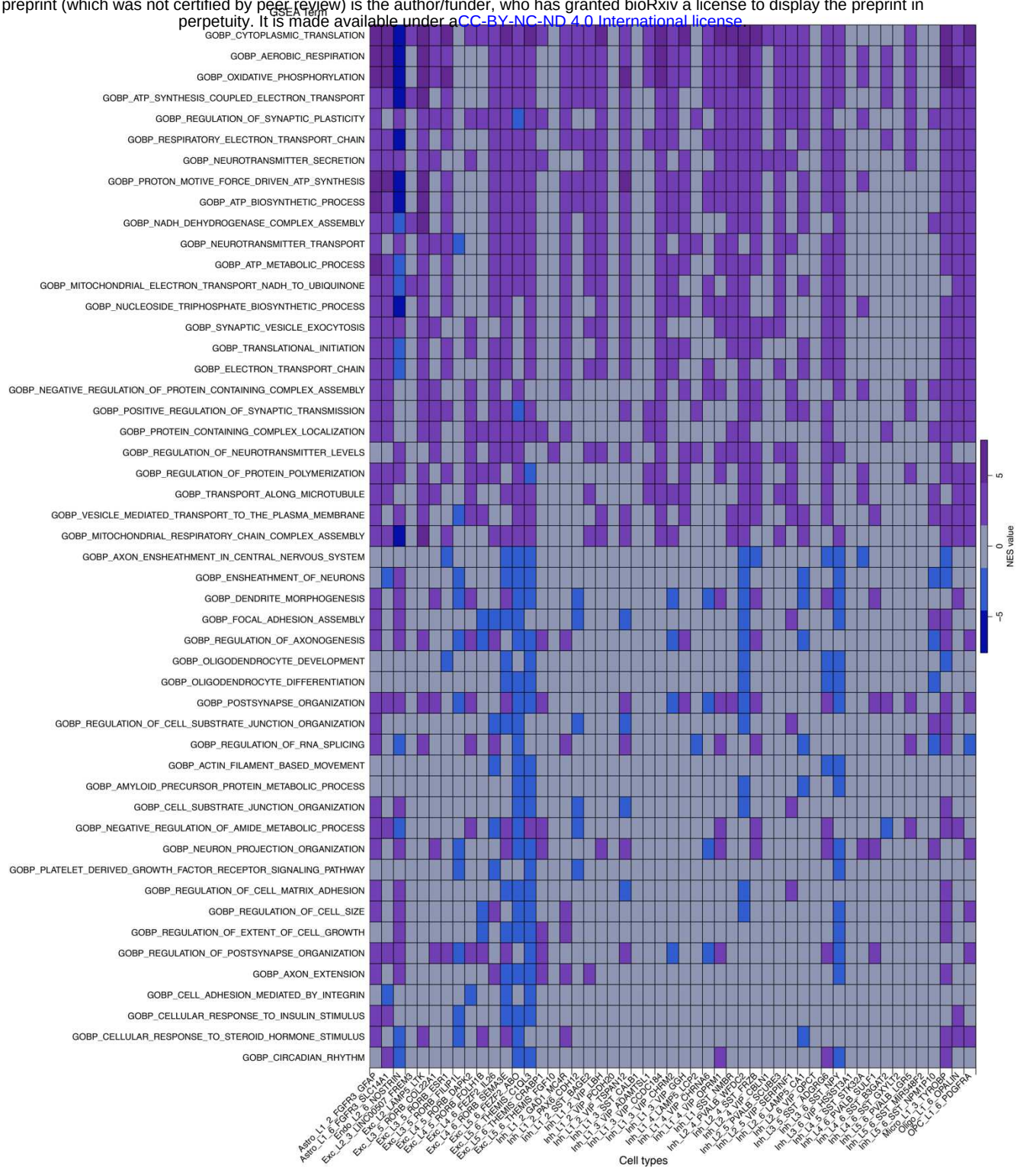
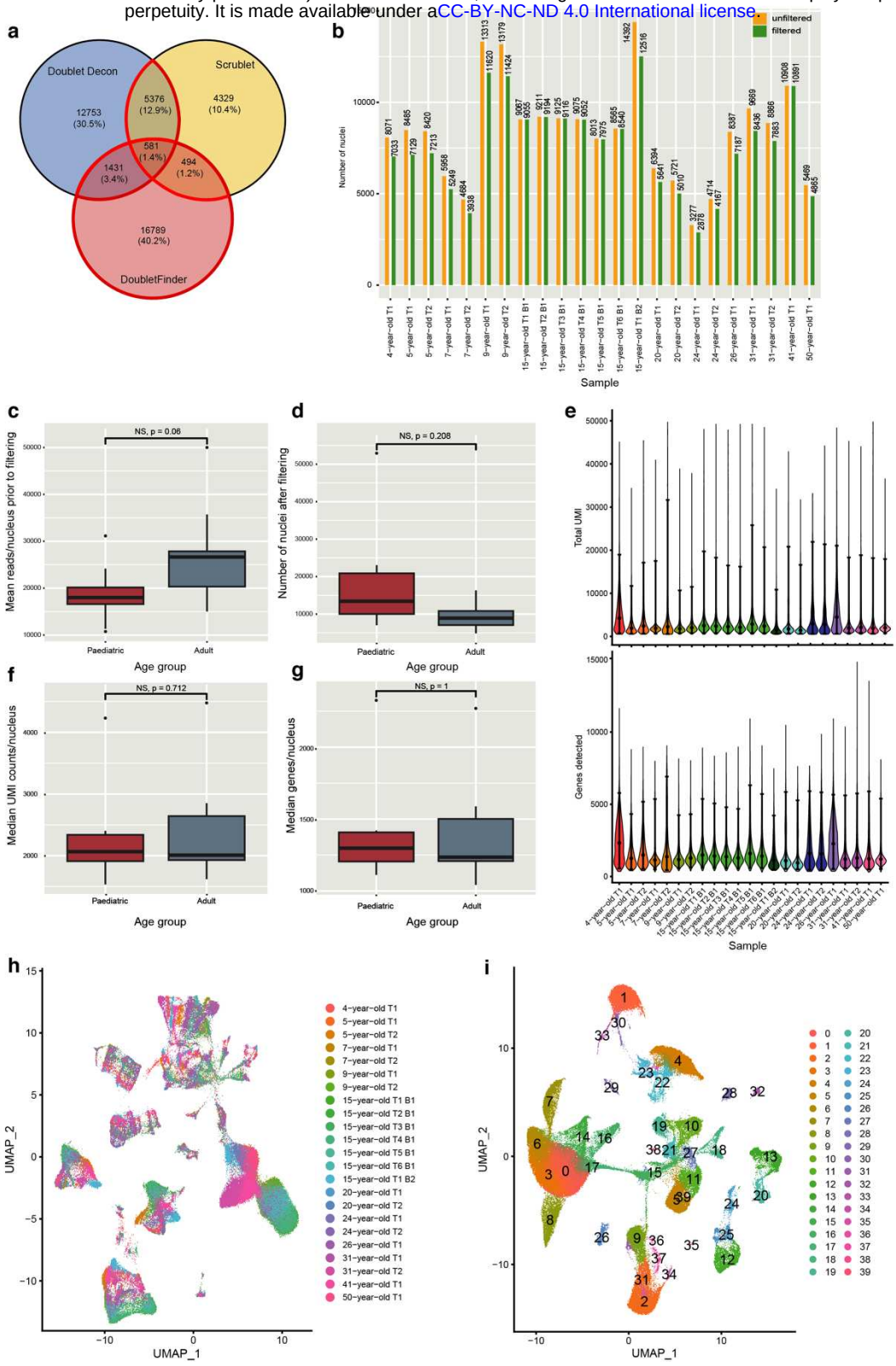
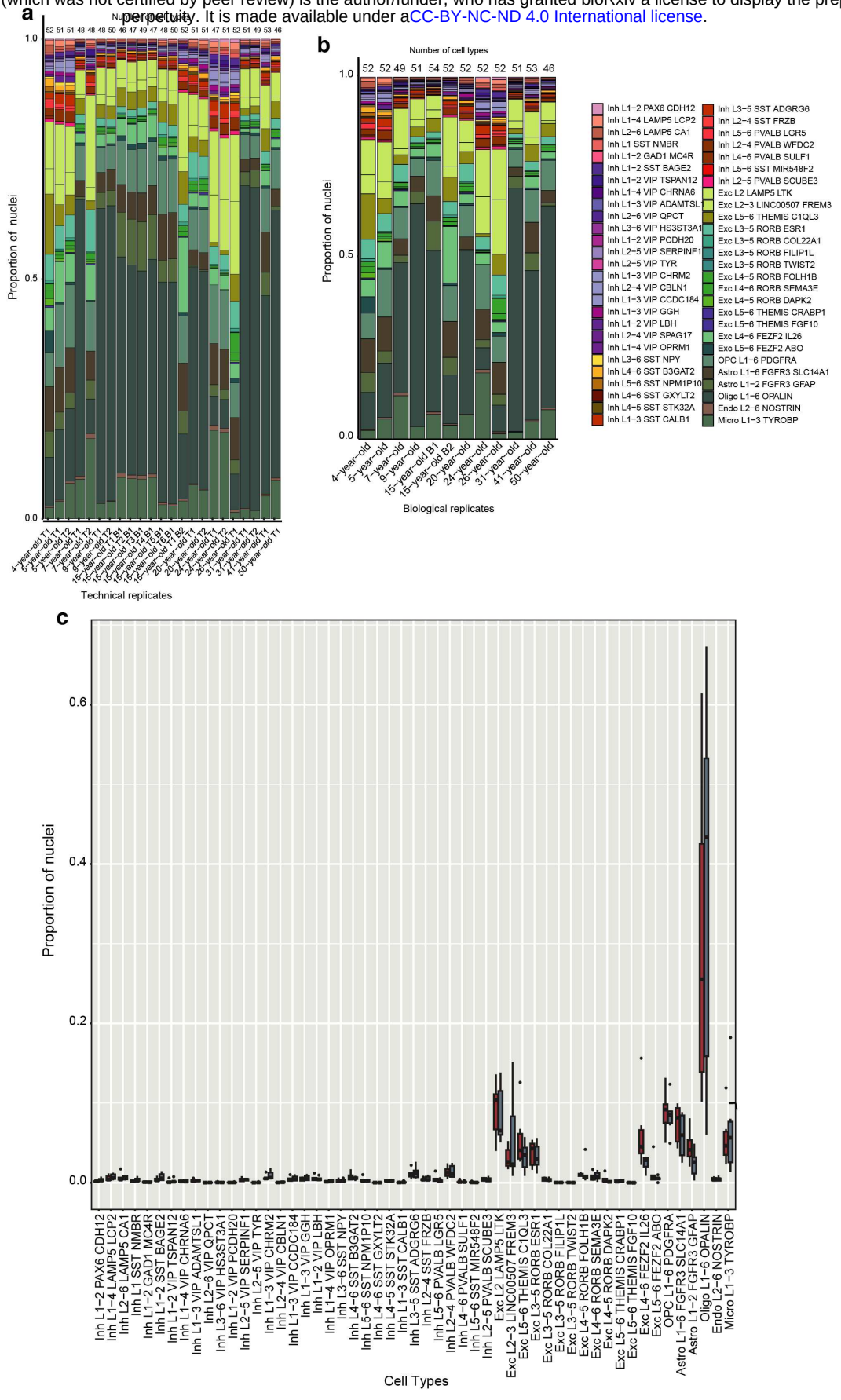


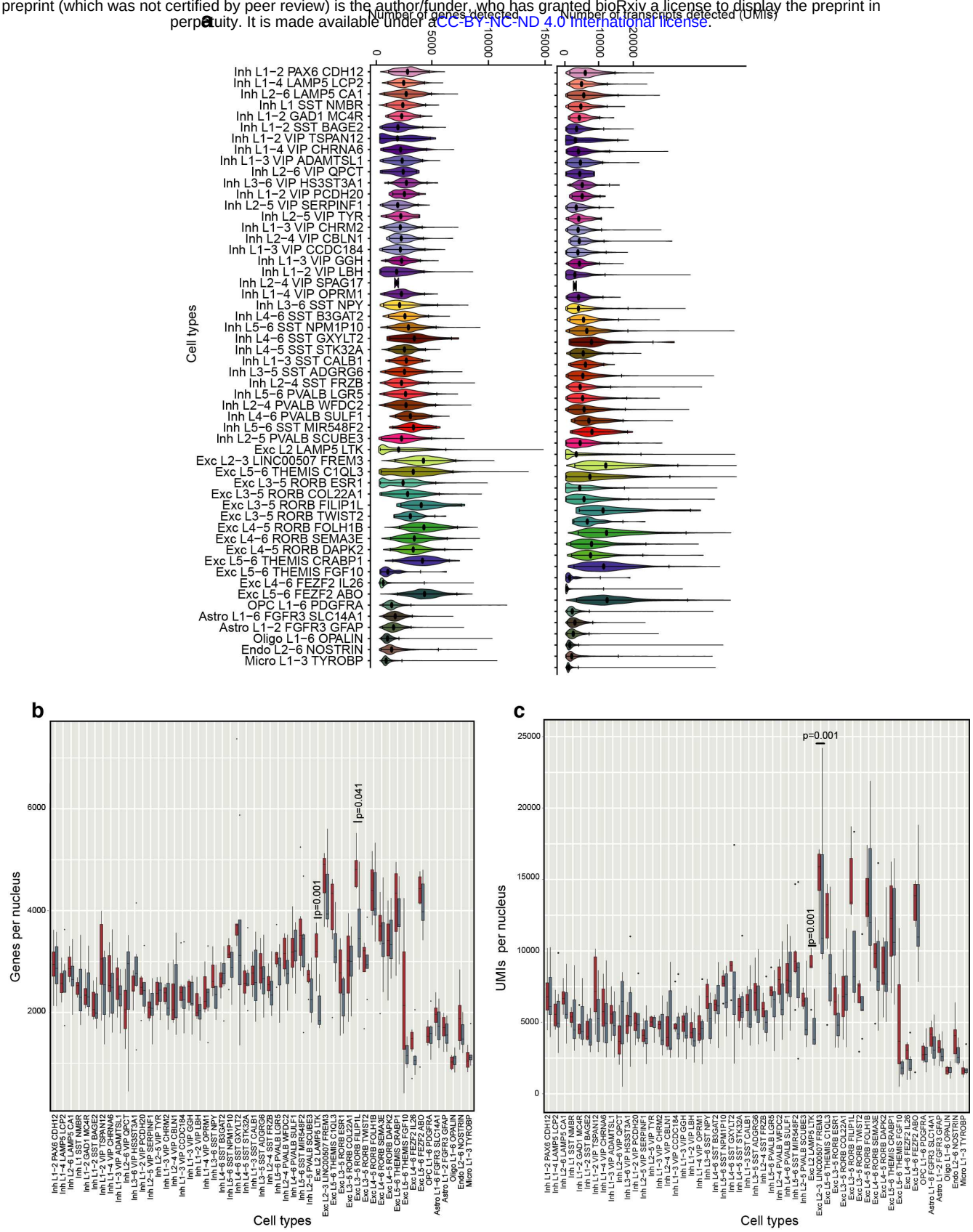
Fig. 6: Pathways that are enriched or depleted across multiple cell types in paediatric cell types. GSEA heatmap showing the top 25 most frequently enriched (top 25 rows) or depleted (bottom 25 rows) terms appearing across all cell types. Only significantly ($p < 0.01$ and $q < 0.1$) terms are shown. NES value represents the normalized enrichment scores. Grey indicates that the term was not significantly enriched or depleted in the indicted cell type. See also Extended Data Table 11.



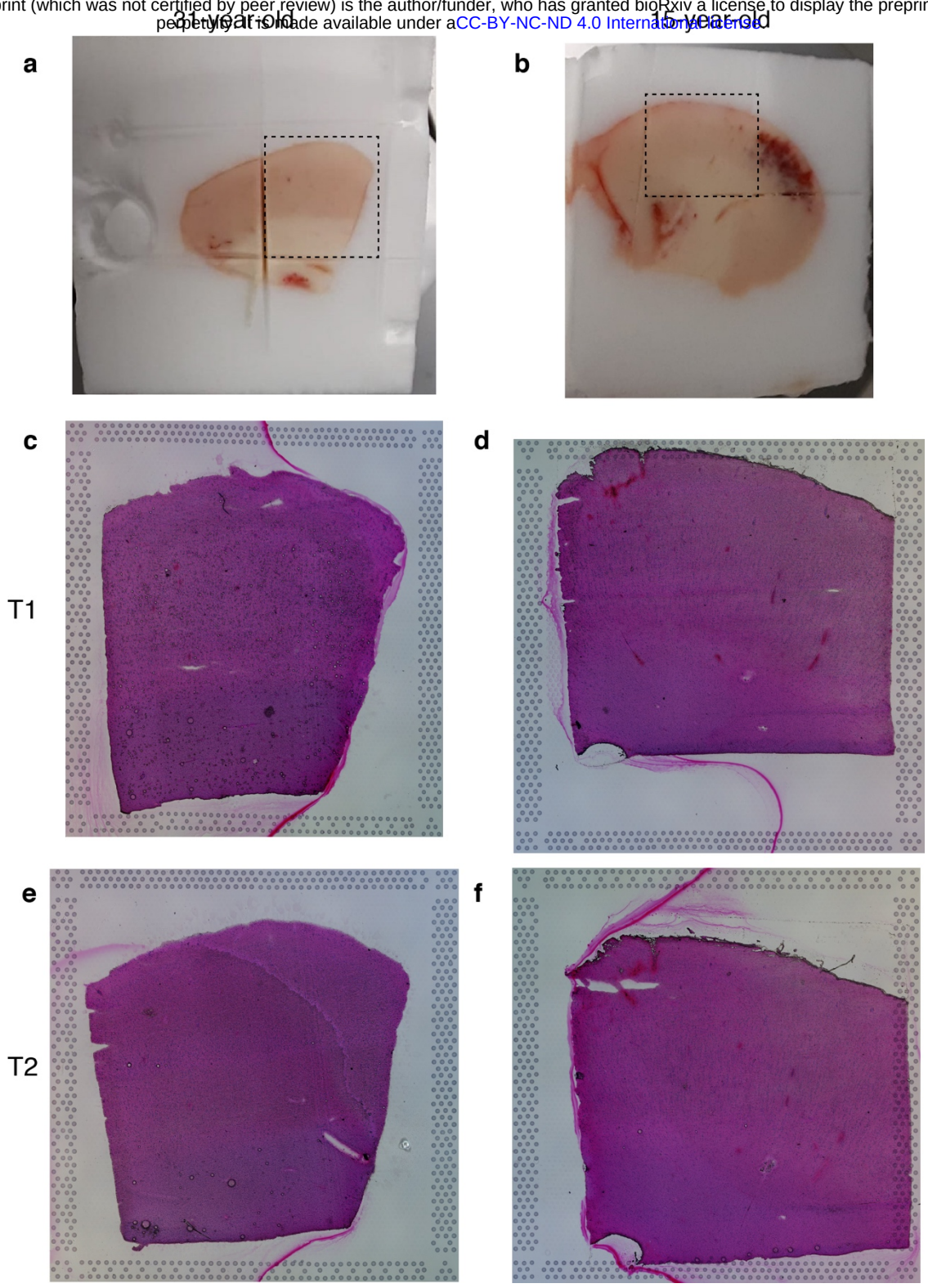
Extended Data Fig. 1: Nuclei filtering and clustering. a, Number of doublets identified across all 23 datasets by DoubletDecon, DoubletFinder, and Scrublet. Red outline indicates the subset of barcodes called as doublets that were removed. **b**, Total number of nuclei per dataset before (yellow) and after filtering (green). **c**, Mean number of reads per nucleus (y axis) by dataset prior to filtering split by age group (x axis). p value determined by two-tailed Welch's t-test. **d**, Number of nuclei (y axis) by sample after filtering split by age group (x axis). p value determined by Brunnermunzel permutation test**. e**, Violin plots showing the number of unique molecular identifiers (UMIs) (top) and the number of genes detected (bottom) per nucleus per sample after filtering. Black dots indicate the median value. Error bars show 95% confidence intervals. **f**,**g**, Median number of UMIs (2,263 paediatric and 2,011 adult) (**f**) and the median number of genes (1,372 paediatric and 1,226 adult) (**g**) detected per nucleus (y axes) by sample after filtering split by age group (x axis). p values determined by two-tailed Brunnermunzel permutation test. **h**, UMAP plot for the 23 datasets prior to integration. **i**, UMAP plot showing the resulting clusters determined by the shared nearest neighbour algorithm. Data in all box plots represent mean ± sem for six paediatric and six adult samples. No significant differences were detected between paediatric and adult samples. B, biological replicate; NS, not significant; T, technical replicate. See also Extended Data Table 2.



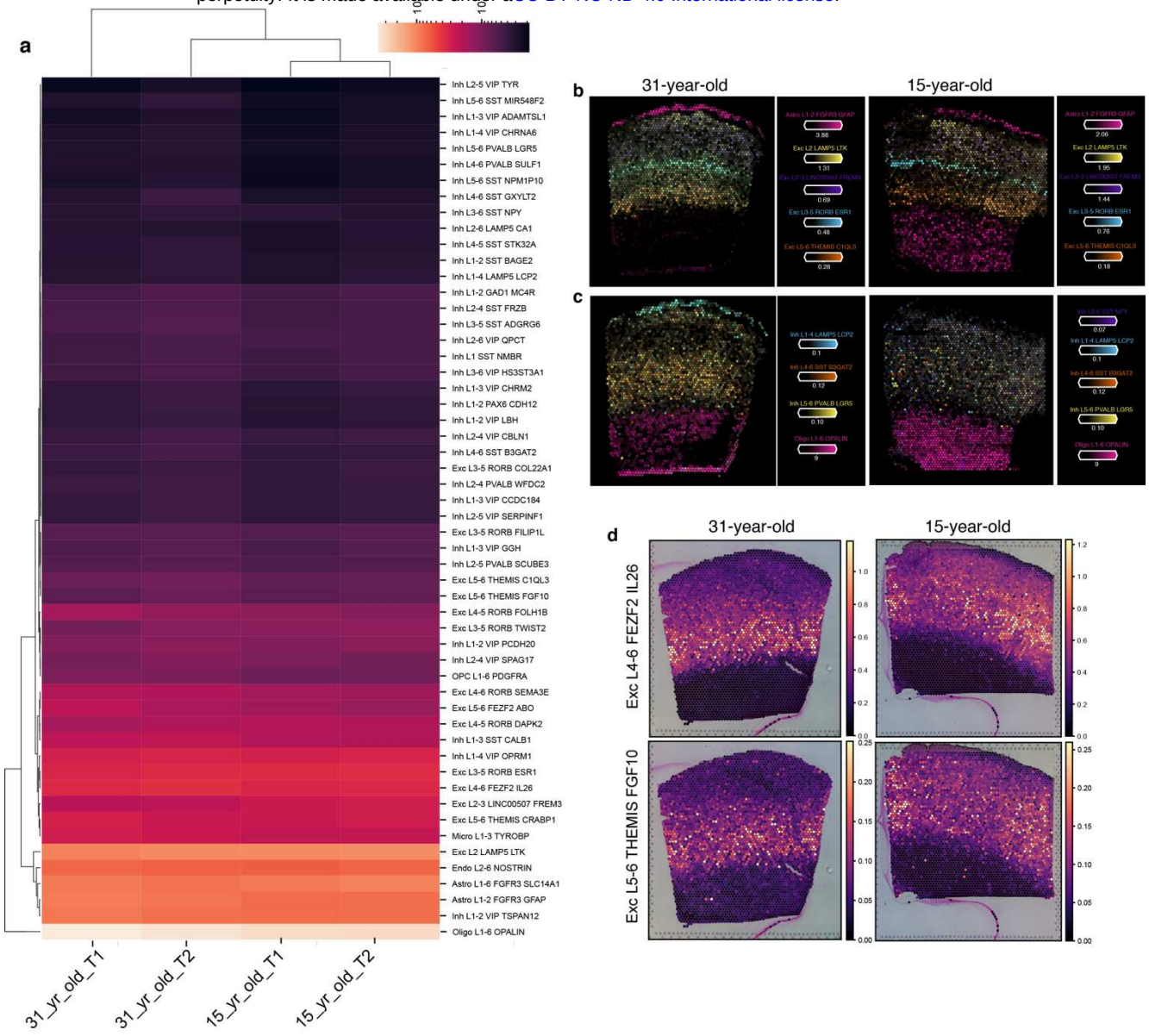
Extended Data Fig. 2: Assessment of cell composition across datasets. a, **b**, Stacked barplots showing the proportion of nuclei per cell type (y axis) for each technical replicate (**a**) or biological replicate (**b**) (x axis) out of the total number of nuclei for each group. The colour scheme for the cell types is in accordance with the MTG cell taxonomy. Samples with technical replicates showed high degrees of similarity in cell composition between their replicates (**a**). Technical replicates from each donor were merged to allow comparisons between the 12 samples (**b**). **c**, Boxplots showing the proportion of nuclei (y axis) per cell type per sample (x axis) split by age group (red: paediatric, grey: adult). Data represents mean ± sem for six paediatric and six adult samples. No significant differences were detected between paediatric and adult samples; See also Extended Data Table 3 for details of statistical tests performed.



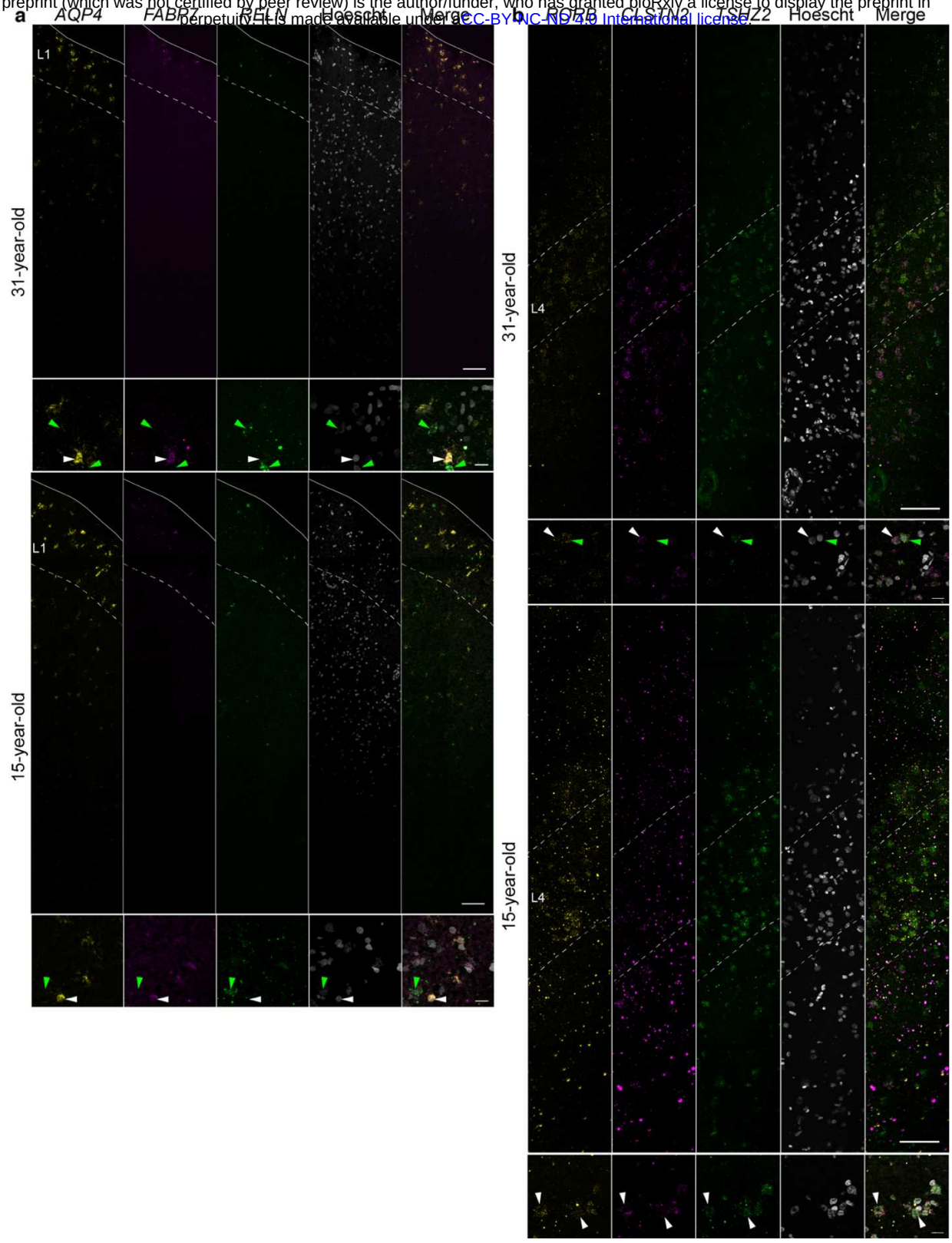
Extended Data Fig. 3: Assessment of the sequencing metrics for the annotated cell types. a, Violin plots showing the distribution of the number of genes (left) and transcripts (right) detected per nucleus per cell type across all datasets. Black dots indicate the median value. Error bars show 95% confidence intervals. **b**,**c**, Boxplots showing the number of genes (**b**) and the number of UMIs (**c**) (y axis) detected per cell type per sample (x axis) split by age group (red: paediatric, grey: adult). Data in all box plots represent mean ± sem for six paediatric and six adult samples for each cell type. p values are given for cell types that showed a significant difference between paediatric and adult samples. See also Extended Data Table 3 for details of statistical tests performed.



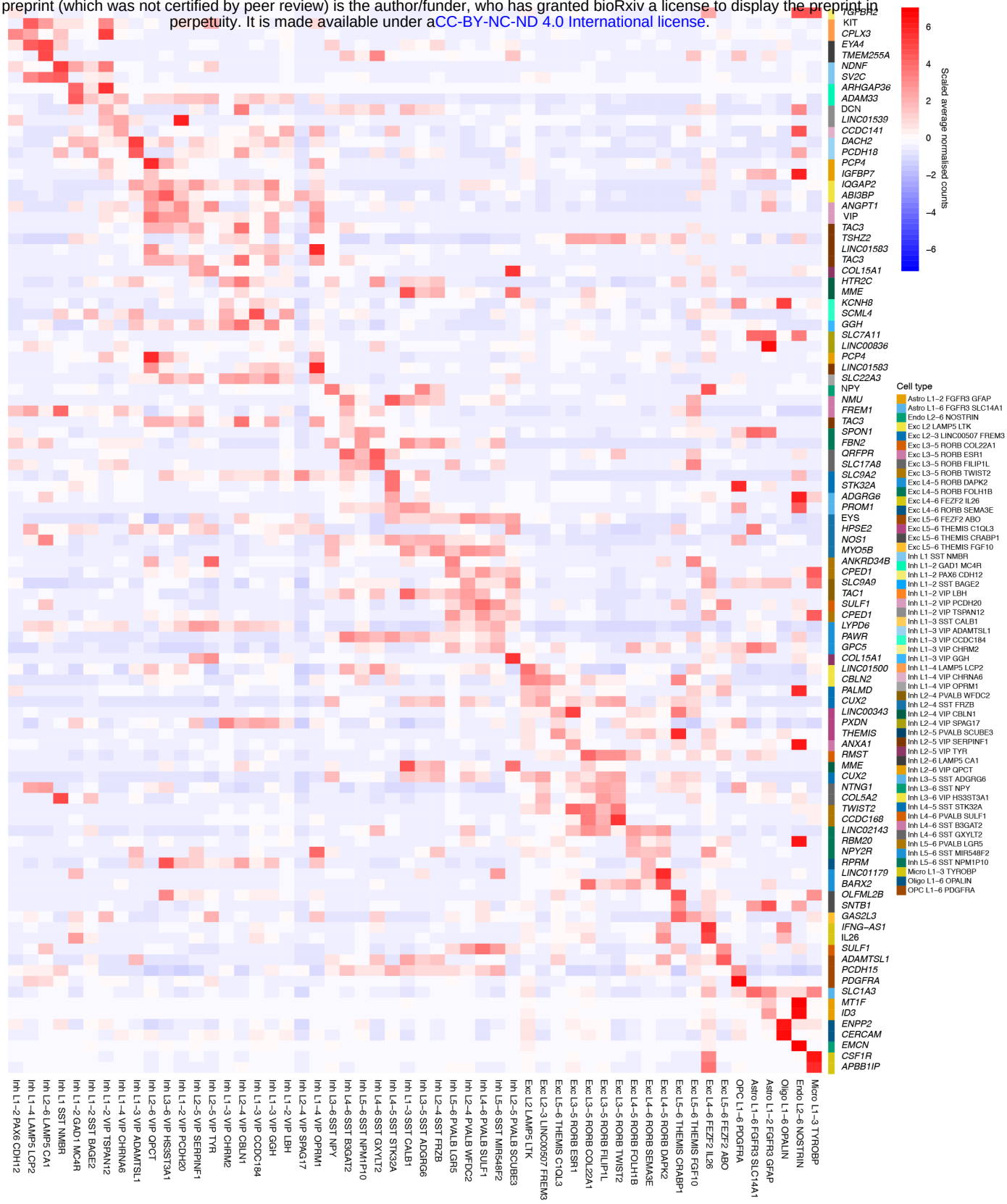
Extended Data Fig. 4: Visium Spatial Gene Expression samples. a,**b**, 31-year-old (**a**) and 15-year-old (**b**) temporal cortex tissue blocks embedded in OCT. Black dashed boxes outline the regions collected onto the Visium Spatial Gene Expression slide. **c**-**f**, H&E stained technical replicate tissue sections used to generate Visium Spatial Gene Expression libraries for the 31-year-old (**c**,**e**) and 15-year-old (**d**,**f**) tissue samples. T, technical replicate.



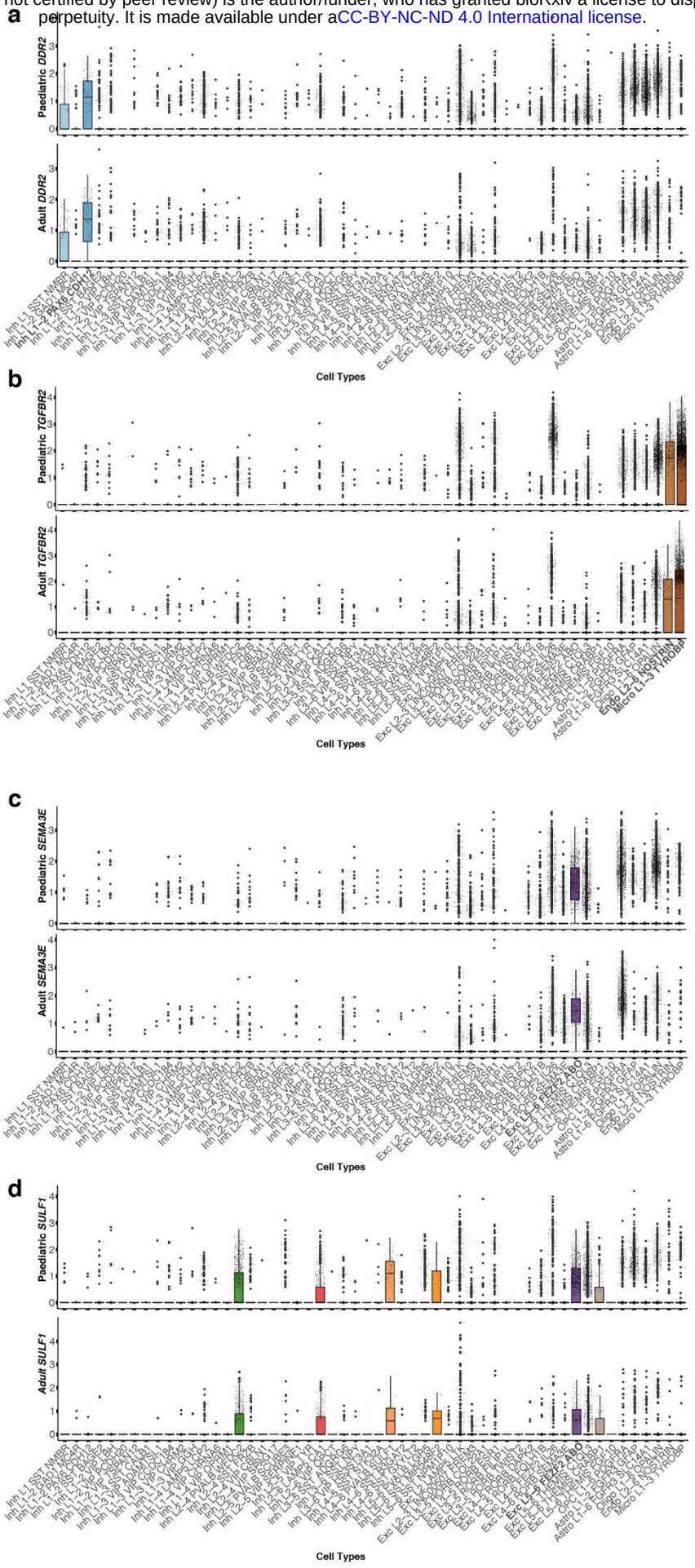
Extended Data Fig. 5: Spatial mapping of cell types in the human temporal cortex. a, Estimated cell abundance of 52 cell types across all Visium samples. Shown is a heatmap with the colour indicating the relative cell abundance of cell types (rows) across the different samples (columns). **b**, Estimated cell type abundances (colour intensity) in the technical replicate 31-year-old and 15-year-old temporal cortex tissue sections for a selection of cell types including non-neuronal cell types (**b**,**c**), excitatory neurons (**b**) and inhibitory neurons (**c**). **d**, Estimated cell type abundances (colour intensity) for Exc_L4-6 _FEZF2_IL26 and Exc_L5-6_THEMIS_FGF10 in the 31-year-old and 15-year-old temporal cortex tissue sections. T, technical replicate.



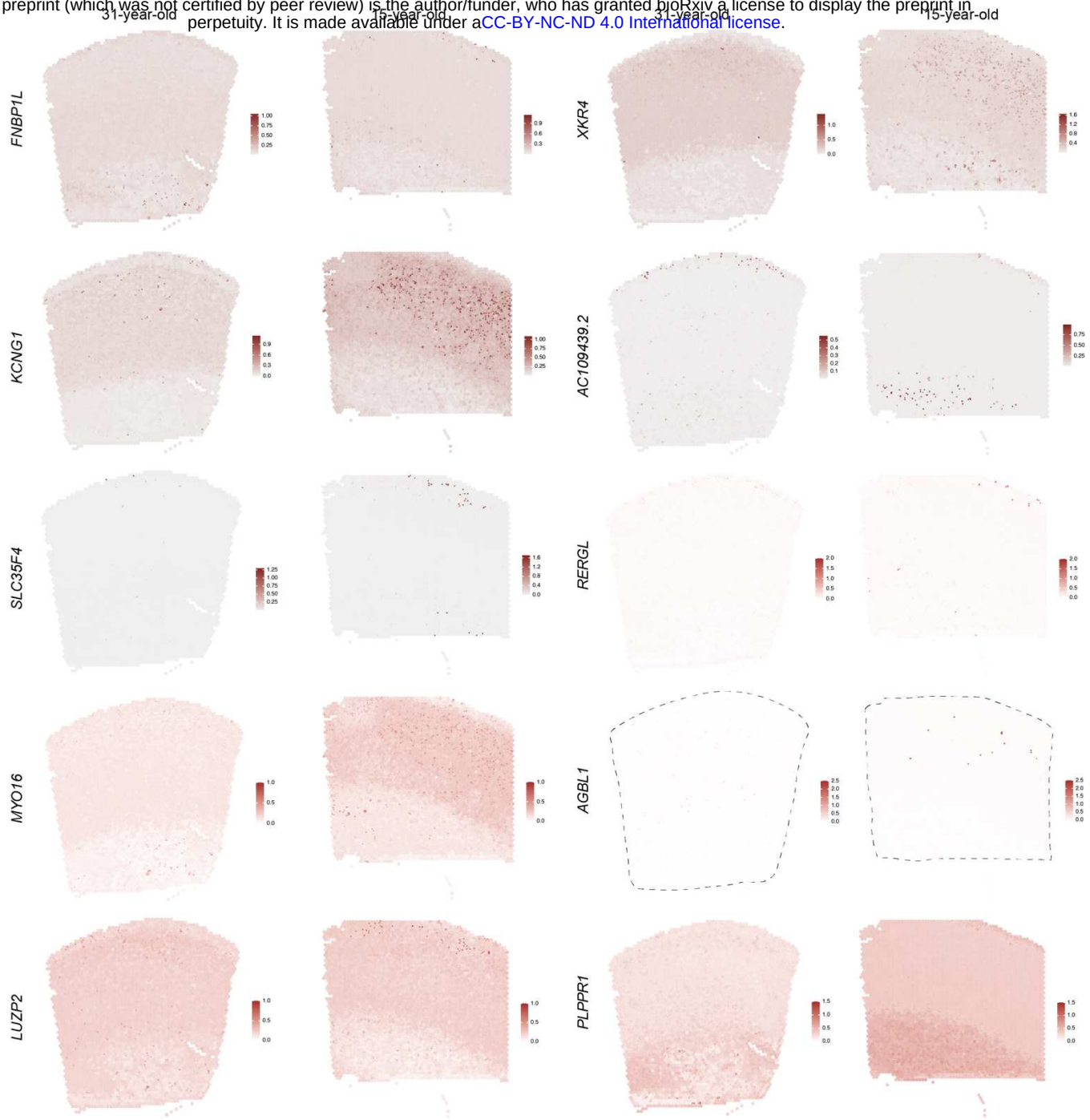
Extended Data Fig. 6: *In situ* HCR analysis of selected cortical layer marker genes. a, Expression of layer 1 markers *AQP4*, *FABP7* and *RELN* in 31-year-old and 15-year-old temporal cortex tissue sections, with high magnification views of layer 1 indicating *AQP4*/*RELN*-positive cells (yellow arrowheads) and *FABP7* positive cells (green arrowhead). **b**, Expression of layer 4-6 markers *RORB*, *CLSTN2* and *TSHZ2* in 31-year-old and 15-year-old temporal cortex tissue sections. In high magnification views of layer 4 in the 31-year-old tissue section *RORB*/*CLSTN2*-positive (white arrowhead) and *RORB*/*TSHZ2*-positive cells (green arrowhead) are indicated. In high magnification views of layer 4 in the 15-year-old tissue section *RORB*/*CLSTN2*/*TSHZ2*-positive cells (white arrowheads) are indicated. Dashed white lines indicate layer boundaries. Solid white line indicates tissue edge. Scale bars are 100 µm in low magnification views (tile scan at 40x) and 20 µm in high magnification views (63x).



Extended Data Fig. 7: Expression of the reference MTG atlas minimal markers. Heatmap showing the scaled average normalised expression counts of the NS-Forest minimal marker genes identified for the reference MTG cell atlas dataset (y-axis) in each of the 54 query cortical cell types identified in the combined adult and paediatric snRNA-seq datasets (x-axis). The minimal marker genes are annotated (colour codes on the y-axes) according to the cell type they define.



Extended Data Fig. 8: Evaluation of NS-Forest minimal marker gene expression across cell types in comparison to MTG cell taxonomy markers. a-d, Boxplots showing the normalised expression counts for *DDR2* (**a**), *TGFBR2* (**b**), *SEMA3E* (**c**) and *SULF1* (**d**) in paediatric (top) and adult (bottom) datasets. The cell types expressing the markers at high levels are indicated in bold.



Extended Data Fig. 9: 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.

Extended data Tables

Extended Data Table 1: Summary of snRNA-seq and Visium 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).

Extended Data Table 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.

Extended Data Table 3: Label transfer annotation of snRNA-seq datasets using the Allen Brain Map MTG atlas as a reference. Sheet 1, Number of nuclei per MTG cell type per sample. 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 2,** Cell types present in reference MTG dataset which are absent from the query datasets. Label transfer resulted in 21 cell populations that were not annotated in the query datasets but were present in the reference MTG dataset. The cluster size and percentage of the total cell count for each cell type in the reference MTG dataset is also shown. **Sheet 3-5,** p values and the test performed for each cell type when comparing the proportion of nuclei (sheet 3; see Extended Data Fig. 2c), number of genes (sheet 4; see Extended Data Fig. 3b) and number of UMIs (sheet 5; see Extended Data Fig. 3c) for each cell type between paediatric and adult samples shown.

Extended Data Table 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.

Extended Data Table 5: NS-Forest minimal marker analysis. Sheet 1, Statistics from NS-Forest analysis including the mean F-beta score (the measure of the discriminative power of a given combination of marker genes) and the mean binary expression score per cell type (a measure of an individual gene's classification power). **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, 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.

Extended Data Table 6: gProfiler analysis of NS-forest markers. (**Sheet1-2**) Significantly enriched GO terms associated with the paediatric (**sheet1**) and adult (**sheet2**) minimal marker genes identified by NS-forest. (**Sheet3-7**) Significantly enriched GO terms associated with shared (i.e associated with both adult and paediatric sample) or paediatric-specific marker genes from the list of top 50 positive NS-Forest markers for selected cell types. Terms for which "highlighted" is true are driver terms.

Extended Data Table 7: Summary of metadata for the top 50 positive NS-Forest markers per cell type. The top 50 positive NS-Forest markers (or total number of positive markers if < 50) per cell type for the down-sampled paediatric and down-sampled adult datasets. The number of intersecting markers, the number of markers unique to paediatric samples, and the number of markers unique to adults 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.

Extended Data Table 8: Overlap of the top 50 positive NS-Forest markers per cell type between the paediatric and adult datasets. The top 50 positive NS-Forest markers (or total number of positive markers if < 50) per cell type were extracted for the down-sampled paediatric and down-sampled adult datasets. Each sheet represents 1 of 53 cortical cell types (Inh_L2-4_VIP_SPAG17 was excluded due to too few nuclei) and the NS-Forest features which were shared between the paediatric and adult datasets, unique to paediatric datasets, or unique to adult datasets are shown (sheet 1-53).

Extended Data Table 9: DESeq2 output of all genes tested for differential expression between paediatric and adult brains per cell type. (**Sheet 1-53**) Differential expression analysis was performed using DESeq2's Wald Test for each cell type separately (Inh L2-4 VIP SPAG17 was excluded due to too few nuclei). 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 log_2FoldChanges 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).

Extended Data Table 10: DESeq2 output of significant DEGs between paediatric and adult brains in a subset of cell types. (**sheet 1-5**) Significant DEGs (padj < 0.05) for Exc_L2_LAMP5_LTK, Exc_L2-3_LINC00507 _FREM3, Exc_L3-5_RORB_ESR1, Exc_L4-5_RORB_DAPK2 and Astro_L1-6_FGFR3_SLC14A1. The associated log_2FoldChanges, 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. Positive log_2FoldChanges 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 6-7**) EA (sheet 6) and IQ (sheet7) associated DEGs and their associated cell types.

Extended Data Table 11: **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 log2FoldChange*-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. 1,290 terms were enriched in at least two cell types, while 250 were depleted in at least two cell types.