Towards a complete human cell atlas: a single-nucleus RNA sequencing study of the paediatric and adult human brain

By Christina Steyn

Master's thesis submitted in fulfilment of the requirements for degree of

Master of Medical Science specialising in Medical Cell Biology

Supervisor: Dr Dorit Hockman

Department of Human Biology

Health Sciences Faculty

University of Cape Town

South Africa

May 2023

Plagiarism declaration

1) I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

(2) I have used the Nature convention for citation and referencing. Each contribution to and quotation in this assignment from the work(s) of other people has been attributed, and has been cited and referenced.

(3) This assignment is my own work.

(4) I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

(5) I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

(6) This thesis has been submitted to the Turnitin module.

Date: 29/05/2023 Signed by candidate

Signed: Christina Steyn

Table of Contents

List of figures

List of tables

Supplementary and Extended Data

Link to Supplementary Data

Supplementary text 1

Supplementary text 2

Supplementary text 3

Supplementary Figure Legends

Supplementary Figure 2.1-2.5

Supplementary Figure 3.1-3.31

Supplementary Table Legends

Supplementary Table 2.1-2.3

Supplementary Table 3.1-3.32

Link To Extended Data

Extended Data Legends

Extended Data 1-5

Code availability

Link to all scripts

Script 1: Cell Ranger pipeline

Script 2: Quality control pipeline

Script 3: DoubletFinder pipeline

Script 4: DoubletDecon pipeline

Script 5: Scrublet pipeline

Script 6: Data normalization, scaling, and integration pipeline

Script 7: Clustering analysis

Script 8: Cluster marker identification

Script 9: SCSA automated annotation

Script 10: scCATCH automated annotation

Script 11: Manual annotation using known cell type-specific markers

Script 12-13: Label transfer using Allen Institute for Brain Science's Smart-seq MTG dataset

Script 14: Broad cell type assignment based on automated and manual annotation methods

Script 15: Validation of MTG cell type annotations by computing similarity scores to compare each of the 54 query cell types to each of the 75 reference cell types

Script 16-18: NS-Forest cell type-specific marker identification

Script 19: Assessing the relevance of NS-Forest marker genes

Script 20: DESeq2 pipeline

Script 21: Psupertime pipeline

Script 22-23: IDEAS pipeline

Script 24: GSEA pipeline

Script 25: Proportion analysis to compare the difference in the percentage of nuclei expressing each gene between paediatric and adult datasets for each MTG cell type

Script 26: Random downsampling to compare the difference in the number of nuclei expressing each gene between paediatric and adult datasets for each MTG cell type

Script 27: Guilt by association analysis

Script 28: Pre-processing Visium datasets

Script 29: Preparing python anndata object from Visium datasets

Script 30-31: Cell2Location pipeline

Abbreviations

snRNA-seq: single nucleus RNA sequencing

scRNA-seq: single cell RNA sequencing

DEG: differentially expressed gene

lncRNA: long non-coding RNA

ASD: Autism Spectrum Disorder

BBB: blood brain barrier

OPCs: oligodendrocyte precursor cells

UMI: unique molecular identifier

NCX: neocortex

DLPFC: dorsolateral prefrontal cortex

QC: quality control

GSEA: Gene Set Enrichment Analysis

GBA: Guilt by association analysis

Mitoratio: mitochondrial-to-normal gene ratio

t-SNE: t-distributed stochastic neighbour embedding

UMAP: uniform manifold approximation projection

GO: Gene Ontology

IDEAS: Individual level Differential Expression Analysis for ScRNA-seq data

OCT: optimal cutting temperature compound

GEMs: gel beads-in-emulsion

LR: likelihood ratio

PCR: polymerase chain reaction

NS-Forest: Necessary and Sufficient Forest

Acknowledgements

To Dorit (my incredible supervisor): I cannot thank you enough for everything you have done for me. Your invaluable feedback on my proposal, various presentations, and thesis has made me a much better scientist and your attention to detail is honestly unparalleled. I am immensely grateful for the opportunity you gave me to spend three months in Oxford and will look back on that time with much fondness for the rest of my life. Thank you for giving me the liberty to explore the data in whichever way I pleased and for guiding me with purpose, patience, and optimism. Your ability to stay calm and detached when experiments fail is truly inspirational. You have taught me that when things go wrong, it is sometimes best to shrug with a smile and simply try again.

To my funders, the Harry Crossley Foundation, the National Research Foundation, the Oppenheimer Memorial Trust, and the UCT Vice Chancellor's Research Scholarship: thank you for having sufficient confidence in me to fund me over the course of my degree. This degree would not have been possible without your unbelievable generosity and support. Moreover, there were times when I have felt doubtful about my own abilities but the knowledge that you believed in the importance of my work and in me as a scientist kept me going.

To my mentors, especially Joe Raimondo, thank you for being so excited about science, for making me feel proud of my work, and for showing me that even great scientists are just humans. To my family: what would life mean without you? Mom, Dad, Emily, and Liam, your encouragement and interest in my work over the past two years means the world to me. Thank you for making me laugh when I thought I was going to cry and for all the little reminders of how incredibly fortunate I am. You have kept me grounded. Tess, my fellow neuroscientist, partner in crime, flat mate, and twin. Thank you for listening (or pretending to listen) to all my rants over the years, for fiercely defending me whenever I was upset, for cooking me so many delicious meals, and for being my personal hype man. To my friends: thank you for keeping me balanced and reminding me to go outside. You all fill my heart with so much joy. Daniel, thank you for staying calm though my many storms, for literally travelling across the world for me, for letting me live with you during the last push to complete my thesis, and for having an unending well of humour. You are amazing.

Lastly, to Science. The Covid-19 pandemic has highlighted your importance to humanity. You are a constant reminder to me that goodness prevails and that there is hope. The scientific process is beautiful in its humility, quest for knowledge, and endeavour to make the world a better place. Working on this project, I have come to love my datasets dearly and feel genuinely excited about their usefulness to the scientific and medical community.

Abstract

Postnatal human brain maturation from birth to early adulthood represents a period of susceptibility for neuropsychiatric risk. While temporal gene expression dynamics over this period have been studied extensively, there are no studies exploring the paediatric brain at single cell resolution. To address this, we present the first paediatric brain cell atlas comprising of 6 single nucleus RNA sequencing (snRNA-seq) datasets generated from ante-mortem human brain tissue samples obtained during elective surgeries to treat epilepsy. To complement these, we included 6 snRNA-seq datasets from adult brain tissue. The 12 samples are all of temporal cortex origin and were produced using the 10X Genomics Single Cell 3' gene expression analysis kits. The datasets were processed using an optimised pipeline and the nuclei were annotated into various cell types using the Allen Institute's middle temporal gyrus dataset as a reference. A novel machine learning method was applied to the annotated datasets to identify combinations of marker genes capable of distinguishing each cell type. Based on this, several minimal marker genes were identified which were shared between paediatric samples and not adults or vice versa. Three different tools were used to identify genes changing in their level of expression with age within each cell type. This revealed hundreds of differentially expressed genes (DEGs), with numerous DEGs being unique to specific cell types and subtypes. From these analyses, two long non-coding RNAs of interest were selected for further *in silico* characterization which revealed putative functions for these genes. Overall, we have provided a resource which can be interrogated further to explore differences between paediatric and adult samples at the gene expression and cell level. This may promote an expansion in our understanding of brain maturation and brain diseases.

Chapter 1: Introduction

1.1. Human brain maturation: an exquisite work in progress

It goes without saying yet serves as an important reminder among paediatric healthcare practitioners that children are not merely small adults. Indeed, this observation is easily recognisable if one examines the everyday behaviour of a child – playful and unfiltered – which often stands in contrast to that of an adult. Less easily discerned however are the underlying mechanisms in the central nervous system governing the wealth of differences that distinguish children from adults. Nonetheless, owing to over half a century's worth of research, a developing picture of the maturing human brain has emerged which captures the phenotypic characteristics and events that define the brain at various stages of postnatal development[1,2]. To begin to understand this picture, it is useful to construct a hierarchy describing how information is organised in the brain (Fig 1.1).

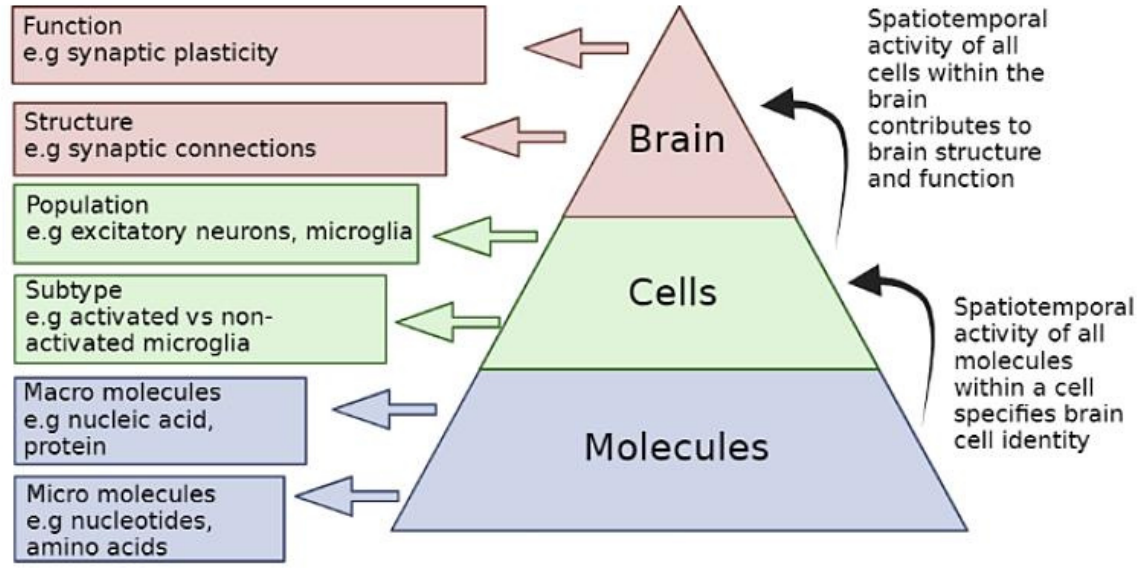


Figure 1.1. The brain as a hierarchy of information. The precise spatiotemporal regulation of micro and macromolecules gives rise to various cell types, subtypes, and cell states, which are themselves regulated in time and space to give rise to the structural and functional patterning of the brain.

The brain is comprised of cells (approximately 171 billion[3]), which are themselves comprised of an outstanding diversity of molecules. Ultimately, brain functions such as synaptic plasticity are underpinned by brain structure such as synaptic connections. For brain structure to arise, various populations of diverse cell types must interact, with each broad cell type usually comprising an array of cell subtypes or substates such as layer 1 excitatory neurons or activated microglia. While the identities of these cells have historically been defined by their varied morphologies[4], it is now well established that any given brain cell can be more accurately classified according to the set of molecules present

at a particular time in conjunction with other properties such as electrophysiology and morphology[5–7]. Thus, it can be thought of that the spatiotemporal activity of all molecules in a cell specifies that cell's identity and the spatiotemporal activity of all cells gives rise to brain structure and function.

The brain begins to develop at just 20 days post conception[2], starting with the formation of the neural tube which gives rise to the brain, spinal column, and the ventricular system of the brain, a structure required for the supply of nutrients and removal of waste, to and from the brain, respectively[8]. By the end of gestational week 8, the primary foundation of the brain has been laid down with the head of the embryo divided into five subdivisions along the caudal-rostral axis including the three broad segments: the prosencephalon (primitive forebrain), the mesencephalon (precursor of the midbrain structures), and the rhombencephalon (primitive hindbrain)[9]. Notably, neuron production starts on embryonic day 42 and proceeds exponentially such that by mid-gestation it is almost complete with billions of neurons generated from dividing progenitors[10,11]. Neuronal cells form the basic computational units of the brain and are ultimately responsible for every thought, feeling, movement, and experience (conscious and unconscious) we have had and ever will have. The broad categories of neurons include excitatory neurons and inhibitory neurons[12]. Functionally speaking, these can be distinguished by their ability to either promote or prevent the generation of an electrical impulse (action potential) in downstream receiving neurons[13]. This alternative activity of excitatory and inhibitory neurons depends on both the nature of the signalling molecules (neurotransmitters) released from neurons as well as the type of receptors these neurotransmitters bind to on the receiving neurons[13].

Neurons are generated from neural progenitors located at the ventricular zone, a region which eventually becomes the epithelial lining of the ventricles of the brain[2]. In order to form functional circuits, newly generated neurons must migrate from their region of origin and either forge the generation of a new neural pathway or integrate into an already-existing network[14] (Fig 1.2). In the neocortex – the region of the brain responsible for higher cognitive functions such as perception, decision-making, and language – the process of migration and integration culminates in a characteristic organization of cells into six distinct layers[15]. Unlike other regions of the developing brain, migration in the neocortex is inverted with the youngest neurons migrating the furthest distance to the more superficial layers and the older neurons forming part of the deeper layers[15]. Integration is achieved through the elongation of neuronal axons and the arborization of dendritic processes allowing synaptic connections to form between cells[16–18]. Although dendritic branching begins in the first trimester[19,20], it is comparatively slow over this period and maximal branching activity is only achieved during the first or second year, post term depending on the region and layer of the brain[19,21]. Thereafter, dendritic development gradually decreases and is complete at approximately 5-7 years of age[22,23] though there is evidence that this process continues into early adulthood for layer 3 pyramidal neurons[20].

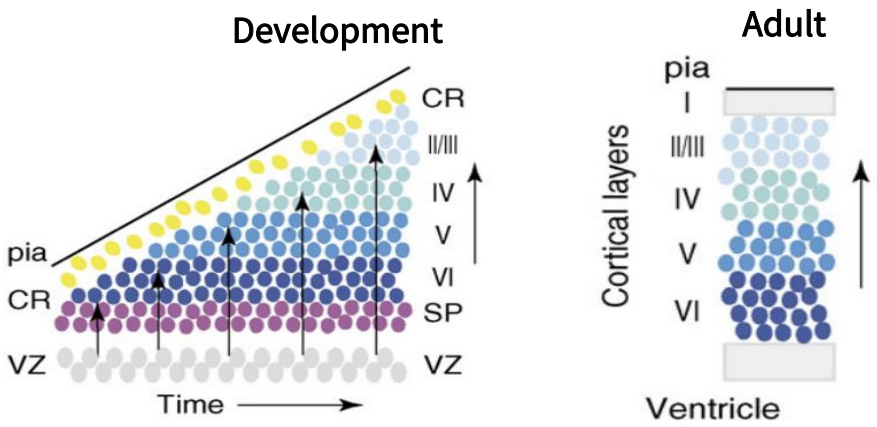


Figure 1.2. Inside-out migration in the developing neocortex. During development, neural progenitors at the ventricular zone give rise to neurons which migrate outward in a temporally regulated manner (horizontal arrow represent the time that new neurons are born). The oldest neurons migrate the shortest distance whilst the newest neurons migrate the furthest distance (vertical arrows represent the distance migrated). This produces six largely distinct cortical layers in the adult neocortex (layers I-VI). The shade of blue indicates the age of the neurons with darker shades indicating older neurons and lighter shades indicating younger neurons. VZ: ventricular zone (grey); SP: subplate cells (purple); CR: Cajal-Retzius cells (yellow); cortical layer cells (blue). Pia refers to the innermost layer of the meninges which are membranes surrounding the brain and spinal cord. Figure adapted from Cooper et al. (2008)[15].

As the brain expands, it begins to form distinctive folds known as sulci and gyri which allows for an increased cortical surface area to fit within the volume constraints of the cranium[24]. This process known as gyrification continues beyond gestation into postnatal development. Additionally, the proliferation and migration of glial progenitors, which begins at approximately 19 gestational weeks, also continue postnatally while the proliferation and migration of neuronal populations are largely complete prior to birth. Neural stem cells known as radial glia (RG) give rise to neurons followed by astrocytes and subsequently oligodendrocytes – with a so-called "gliogenic switch" mediating the transition between neuron and macroglial cell production from RG cells[25]. Astrocytes and oligodendrocyte populations play crucial roles as supporting cells to neurons in the brain. While the diverse array of astrocytic functions is beyond the scope of this review, several key functions include regulating the composition of the extracellular matrix, maintaining the integrity of the blood brain, and removing neurotransmitters from the synaptic cleft[26].

On the other hand, oligodendrocytes function in the production of myelin which is a fatty substance that is deposited over neuronal axons and serves to insulate the axons in much the same way that plastic covers an electrical wire thereby protecting it and increasing the velocity of electrical transmission[27]. The process of myelination is a dynamic one which sees oligodendrocytes supplying myelin to neurons in accordance with their requirements[27,28]. There is evidence that this process continues beyond the second decade of life for some cortical fibres[29] and possibly throughout the lifespan[30], though peak myelination occurs in the first year of life corresponding with heightened axonal sprouting over this period[1,28]. Babies are born with the vast majority of axons unmyelinated, and it is proposed that the increase in myelination that occurs during childhood contributes to enhances in cognitive

functions such as information processing speed as the brain matures[31]. The predecessor to the more mature oligodendrocyte population is the oligodendrocyte precursor cell population which has increasingly been implicated in functions besides serving as a progenitor population, including intercellular signalling with neurons[32,33] and other glia[34,35] as well as protecting unmyelinated axons by extending their processes around them[36]. In support of the broader functionality of OPCs is the observation that they frequently make connections with other cells[32,34,35]. Additionally, they are distributed throughout both the grey and white matter unlike oligodendrocytes which are largely located in the white matter[37]. Overall, there exists convincing evidence for the necessity of macroglial cells in mediating the structural and functional maturation of neuronal circuits.

A third glial cell population is the microglia population which are derived from primitive myeloid progenitors of the yolk-sac and infiltrate the CNS during early foetal development prior to astrogenesis and oligodendrogenesis[38]. They are best known for their roles in detecting and removing pathogens from the brain as well as phagocytosing cellular debris such as protein aggregates and dead cells which can damage the CNS – thus emulating the functions of macrophages in the rest of the body[39]. When challenged by either injury or neuroinfection, microglia are capable of launching either pro- or anti-inflammatory responses depending on a complex interplay of events in the microenvironment which is still not fully understood[40,41]. In more recent years, microglia have also been implicated in synaptic pruning[42] which is an essential postnatal process required for the refinement of neuronal circuits. The outcome of synaptic pruning is an increase in the precision of the more coarse connections initially generated, thereby making functional neural pathways more effective[42]. A group of researchers led by Cornelius Gross demonstrated in two seminal papers that microglia interact with excitatory neurons and essentially nibble away at their synaptic processes contributing to a loss of connections between cells[42,43]. Without this process of synaptic pruning, there would be an excess of synapses which could have negative implications for learning ability[44]. Intriguingly, this overabundance of synapses is a phenotype which has been observed in autism spectrum disorder (ASD) while an opposing phenotype of diminished synapses has been found in Schizophrenia[45,46]. In addition to synaptic pruning, the programmed cell death of approximately 50% of neurons is a similar seemingly regressive but controlled procedure which occurs over the perinatal period[47–49]. This is also thought to be part of a refinement strategy to not only fine-tune neuronal networks but also eliminate neurons whose axonal processes were targeted to the wrong region[47].

Regarding the temporal context of synaptic pruning, it is preceded by a period of 'exuberant' synaptogenesis in the early postnatal months[50,51] with as many as one hundred trillion synaptic connections generated by age 2 – approximately 50% higher than the synaptic density of the adult brain[51,52]. Although this excessive formation of synaptic connections appears to be an energetically wasteful process, it may in fact promote more

robust and efficient neural networks in the long term by facilitating the formation of optimal network structures[53]. Pathways which are continuously used are reinforced due to the connections between synapses becoming stronger over time, a mechanism known as long term potentiation, which underpins our ability to learn and form memories[54,55]. On the other hand, those pathways which are not frequently used gradually deteriorate as a result of connections being cut back, a concept which is captured by thenotion, "If you don't use it, you lose it". Where synaptogenesis peaks in the early postnatal period, synaptic pruning activity begins during late childhood, peaks during adolescence, and gradually decreases from then on[52,56,57]. In contrast to the long-held understanding that synaptic pruning activity levels out during early adolescence, Petanjek et al. (2011) provided evidence that this event extends much further into brain maturation than previously thought with a plateau in spine density only being reached in the late twenties[57]. Together, synaptogenesis and synaptic pruning influence the property of the brain known as plasticity which allows the brain to change its structure and function in response to experience[58]. Considering synaptic changes are most prominent in infancy, childhood, and adolescence, these stages represent periods of enhanced susceptibility to environmental influence. This may in part explain the sponge-like ability of children to absorb information as seen by their remarkable capacity to learn language[59], or the reason why one is at increased risk of developing a substance use disorder if an addictive drug is taken before age 15[60]. Arguably however, the brain retains plasticity throughout the lifespan albeit to a lesser degree as one ages[61].

Perhaps one of the most striking differences between prenatal and postnatal brain maturation is the underlying principle by which the maturation of neuronal circuits is governed. Where prenatal circuit formation largely relies on genetically determined developmental programs[62–65], postnatal maturation of neuronal circuits is increasingly driven by stimulus-dependent changes in gene expression[66,67]. Nevertheless, during both prenatal and postnatal periods, brain cells receive and process signals from both intrinsic and extrinsic factors. For example some *in utero* stimuli such as exposure to neurotoxins can profoundly influence foetal brain development despite the changes over this period mostly following a genetic blueprint[68]. Likewise, postnatal circuitry is still shaped to some extent by genetic context resulting in characteristic micro and macroscopic features emerging at distinct developmental stages[57,62,69,70]. However, after birth, a variety of stimuli not previously encountered become available to the brain in the form of visual, auditory, olfactory, gustatory, and tactile inputs[71]. These are received via specialised receptors capable of detecting specific stimuli and converting them into electrical impulses that are sent to the brain for processing[72].

As the brain matures, regions become specialised to process certain inputs and the early years of life represent a critical period for normal structural and functional patterns of brain organization to be established. For example, individuals who are born deaf and are later provided with cochlear implants that allow them to receive auditory inputs, tend to perform

better in speech and hearing tasks if they received the implants prior to age 5[73], suggesting that the first 5 years of life are important for patterning the brain to be able to receive and process auditory stimuli. To demonstrate the necessity of environmental inputs on brain wiring, pioneering researchers in the 1970s showed that when a 2-week-old monkey was deprived of visual inputs in one eye, the visual processing regions of the brain failed to form the characteristic striped patterning of alternative inputs normally observed when stimuli are received from both the left and right eye[74]. Instead, the visual processing area became dominated by inputs from the open eye which formed thicker bands while inputs from the closed eye formed much thinner bands[74]. This visual deprivation in the young animals had long lasting consequences resulting in visual impairments even after visual stimuli was again made available to both eyes, while the equivalent experiment in adult animals did not affect either brain structure or vision as patterning had already been established[67]. Alongside deprivation studies, researchers experimented with damaging the visual cortex during early postnatal development and found that the inputs which would normally be processed in this region invaded other regions of the brain such as the temporal cortex which usually only processes auditory stimuli[75]. This seeming adaptability of neural pathway formation is once again highly restricted to a short window period in early infancy and damage to the visual cortex in adult animals usually results in visual impairment with none of the abovementioned compensatory mechanisms[75].

In addition to deprivation and elimination experiments, researchers performed enrichment experiments where they showed that by simply raising animals in a stimulating environment compared to a standard laboratory cage, they could alter brain structure and function[76–78]. Indeed, those animals reared in complex environments with other littermates and changing scenery had a greater number of synaptic connections, more glial cells, increased myelination of axons, as well as improved cerebrovasculature[76–78]. Markedly, these changes persisted even after the animals were returned to average conditions[77] suggesting that positive early-life experiences may provide individuals with a physiological robustness that can compensate for adversities in later life. Thus, the formative years of postnatal brain maturation represent periods of both vulnerability and adaptability with regards to neuronal connectivity and function. Taken together, the abovementioned studies demonstrate that for normal brain maturation to occur postnatally, environmental inputs are essential. Overall, it is the integration of diverse experiences with intrinsic signals within specific time frames which determines developmental outcomes[79]. Besides the changes occurring at the level of the synapse, postnatal human brain maturation encompasses a multitude of other structural and functional changes. These include alterations in glucose metabolism which correlate with changes in synaptic density in that cerebral energy metabolism increases during infancy, resides at a consistent high level during childhood, and then begins to fall during adolescence[80]. Volumetric changes in cortical grey and white matter have also been observed during brain maturation using MRI

scans, with an overall reduction in brain volume shown to occur with age, however, white matter volumes increased during childhood and adolescence[81]. These white matter increases suggest a maturation of fibre tracts in juvenile brains which was further confirmed by diffusion tensor imaging studies showing that myelination of axons increases between 5 and 12 years of age[82]. For some fibres such as association fibres, myelination continued to progress well into adulthood[83] demonstrating that the brain requires many years of experience to achieve full maturity. In contrast to grey matter *volume* which has been found to decrease with age[84], grey matter *density* (previously believed to be correlated with grey matter volume) increases from childhood to young adulthood which likely coincides with the timing of certain cognitive milestones such as enhancements in attention, working memory, and response inhibition[12,85]. In addition to volumetric and density changes as the brain develops postnatally, there appear to be certain advances in the functioning of the blood brain barrier (BBB), with some studies suggesting that the paediatric BBB is at increased risk of disruption[86]. Nevertheless, functional barrier mechanisms are already developed in early life[86]. The timing of the key postnatal milestones described above are summarised in Figure 1.3. These events exemplify the protracted process of brain maturation from birth, through childhood and adolescence, into early adulthood.

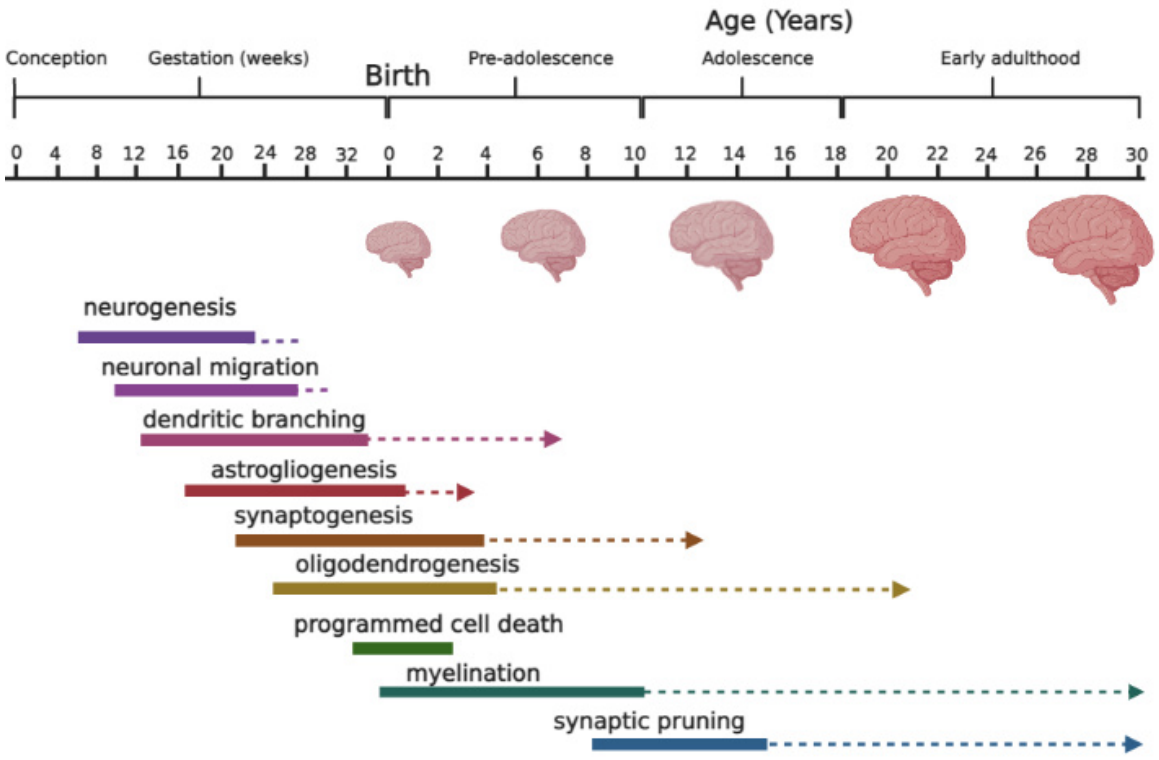


Figure 1.3. Timing of key events during human brain development. Schematic showing the temporal sequence of important developmental milestones occurring over the course of human brain maturation, many of which start before birth and continue for a period postnatally. Processes such as synaptic pruning and myelination are protracted and extend into adulthood. Figure adapted from Tau & Peterson et al. (2010)[12] and Silbereis et al. (2016)[87]

1.1.1. Rationale for studying the molecular and cellular dynamics of the developing postnatal brain

Postnatal brain maturation from birth to early adulthood represents a period of susceptibility in term of neuropsychiatric risk[88] with disrupted trajectories of gene expression associated with neurological disorders such as ASD and Schizophrenia among others being observed over this period[45,46]. While the events defining brain maturation have been well characterised at a broad level, the molecular regulation underlying these changes is less well understood. Describing typical gene expression trajectories as the brain matures postnatally may serve as an important reference to assess the effects of genetic perturbations and early adverse experiences on brain maturation and thereby gain insight into the developmental origins of neurological disorders. Furthermore, investigating the driving forces behind maturational processes may prove useful when it comes to developing interventions for treating and managing neurological disorders since implicated genes may one day serve as therapeutic targets or biomarkers for improved diagnosis and prognosis of conditions[89]. One study which highlights this potential, examined the transcriptomic and epigenomic regulation of the brain across the lifespan and identified numerous genes involved in neurodevelopment which have also been associated with neuropsychiatric conditions[90]. For example, the gene *MEF2C* was found to decrease in expression over the perinatal period and is known to be implicated in synapse function and ASD – thus representing a possible therapeutic candidate for treating adverse symptoms associated with ASD[90].

Considering that many neuropsychiatric disorders show distinct periods of onset with more than half of all conditions being diagnosed by the age of fourteen[91–93], the period of brain maturation under investigation is a crucial period in terms of understanding how these conditions develop. With sufficient metadata and a large enough sample size, this line of research could potentially help to elucidate how childhood experiences influence the genetic regulation of the brain and in turn one's ability to cope with psychological stress in adult life. Moreover, studying the gene expression dynamics of paediatric and adult brains in conjunction may be informative in terms of understanding why the same neurological conditions can manifest differently between children and adults as well as explain differential responses to treatment between the two age groups[94]. Within the South African context this is imperative given the high rates of certain neurological conditions such as paediatric epilepsy and traumatic brain injury due to neuroinfection, accidents, and abuse, with few precision treatments for the paediatric brain[95,96].

1.1.2. Single cell transcriptomics as a strategy to interrogate brain complexity

The human brain has been cited as the most complex object in the known universe by physicists and biologists alike[97,98]. Contributing to this complexity is the diverse array of cell types comprising the brain. As described in the introduction, brain function is ultimately dependent on the coordinated activity of various cell types serving specialised functions ranging from computational roles (neurons) to protective roles (microglia), and supportive roles (astrocytes and oligodendrocytes). For cell type-specific functions to arise, precise spatiotemporal regulation of gene expression is required[87]. The resulting transcriptomic diversity of this carefully controlled system can now be assessed with unprecedented accuracy using single cell transcriptomic technologies which reveal the transcriptomes of individual cells, allowing one to distinguish different cell types and states from one another in an unbiased manner. Taking advantage of this innovative method, the NIH BRAIN Initiative Cell Census Network (BICCN)[99] and Human Cell Atlas Project[100,101] aim to develop a complete picture of the molecular and cellular diversity characterising brain structure and function. Recently, the first major outputs from these cell atlas projects have been published which, among other notable achievements, includes successfully defining brain cell types and subtypes by their distinct transcriptional profiles[102–111].

The single cell studies which have been conducted to date have already yielded key insights including the observation that many glutamatergic neuron subtypes are region-specific in the neocortex while subtypes of GABAergic and non-neuronal cells are usually shared across neocortical areas – a finding which was consistent between two complementary studies of the adult mouse and human cerebral cortex[110,112]. Additionally, Bakken et al. (2021) identified a set of conserved cell type-specific marker genes across humans, marmosets, and mice as well as a set of cell type-specific marker genes unique to human brain cell types. In future, these genes may be interrogated further as putative cell type-specific regulators of brain mechanisms which are essential across species or mechanisms which are human-specific[108]. Importantly, single-cell and spatial transcriptomic technologies are already being applied to the study of neurological disorders, revealing novel cell type-specific or layer-specific dysregulation of gene expression associated with various conditions such as autism[113,114], major depressive disorder[33], and Alzheimer's disease[115].

While most single cell transcriptomic studies of the brain have used postmortem tissue or mouse tissue, there are several studies which have used ante-mortem tissue, including a ground-breaking study by Darmanis et al (2015) who compared the transcriptomes of the adult and fetal brain at single-cell resolution[111]. From this analysis, they identified both known and novel marker genes distinguishing fetal neuronal progenitors, mature fetal neurons, and adult neurons[111]. Furthermore, a subset of adult neurons was found to express major histocompatibility genes contrary to the historical understanding of adult brain neurons being immunologically inert[111]. This discovery may have previously been obscured

due to the use of bulk RNA sequencing methods which are not sensitive to detect gene expression signals within subsets of cells. More recently, Nelson et al. (2022) examined single cell gene expression profiles of 75 individuals using epilepsy and tumor brain tissue samples[116]. Through their analysis, they identified variation in the abundance of different cell types and in gene expression profiles across the samples with factors such as sex, age, and ancestry being contributors to the observed variation. Although these studies are incredibly informative in their own right, they are limited in that they only investigate the single-cell transcriptomic profiles of the fetal and adult brain while the in-between stage regarding the paediatric brain is not addressed[111,116]. The lack of paediatric single-cell datasets is a crucial gap which pertains not only to these studies[111,116] but to the human brain single-cell literature more widely[90,108–110,117,118].

1.1.3. A comparison of bulk RNA sequencing and single-cell RNA sequencing

Before the advent of single-cell transcriptomics, researchers relied on DNA microarrays or bulk RNA sequencing to profile gene expression in the human brain[119–121]. These approaches are limited when compared to the current available single-cell technologies in that the expression signal measured comes from multiple cells within the tissue resulting in an average signal for each gene (Fig 1.4). In complex tissues such as the brain which comprise of multiple cell types, this is especially problematic as one cannot infer whether a given gene is expressed across many cell populations or is expressed in specific cell types only[122]. An initial strategy to address this limitation was to purify individual cell types using population-specific markers and perform RNA sequencing on sorted cell types[123,124]. However, this still results in an averaging of the expression signatures from multiple cells without being able to examine the heterogeneity within a cell type[125,126].

In contrast, single cell RNA sequencing methods (scRNA-seq), as the name suggests, sequence the transcriptomes of individual cells, allowing one to discern the signal specific to each cell. Broadly speaking, this involves separating individual cells and supplying each cell with a unique primer set that contains a molecular barcode specific to that cell such that the molecular barcodes can be sequenced along with the transcripts. As a result, the transcripts can be associated back to their original cell and quantified at the gene-level using computational methods. The 10X Genomics method uses the Chromium Controller microfluidics system which separates cells and captures them individually in partitioning oil along with a gel bead containing a unique primer set (Fig 1.5). This allows all transcripts in a given cell to be converted to cDNA using the same primer set such that they receive a unique DNA molecular barcode that can be used to trace them back to their cell of origin post-sequencing using computational means.

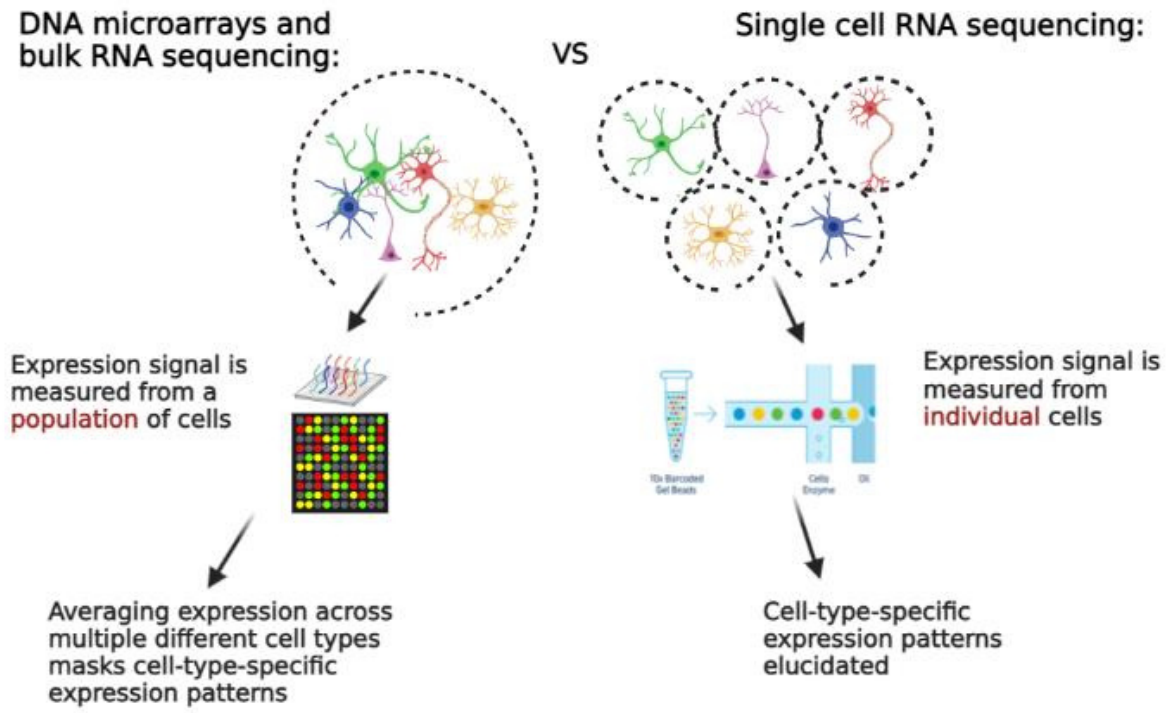


Figure 1.4. A comparison of DNA microarrays and bulk RNA sequencing methods with scRNA-seq methods. Profiling gene expression using DNA microarrays or bulk RNA sequencing makes use of whole tissue samples comprising a population of cells such that the signal measured is an average expression signature from multiple different cell types. scRNA-seq methods profile the transcriptomes of individual cells allowing for cell-type specific expression patterns to be elucidated.

Many single cell transcriptomic protocols also provide each original transcript with a unique molecular identifier (UMI) prior to PCR which functions to minimise amplification bias associated with PCR since all transcripts with the same UMI correspond to a single transcript count (Fig 1.5)[127]. The same approach can be applied to nuclei (snRNA-seq) which is useful for generating human brain tissue datasets since these samples are usually frozen for logistical reasons and thus cells cannot be effectively isolated whereas nuclei can[128,129]. The use of nuclei is also advantageous over cells in that there is less bias in which cell types are captured and there is a reduction in transcriptional artefacts due to the isolation process[128,129].

Mapping gene expression dynamics in the human brain is comparatively challenging to that of other organs and species for several reasons not least of all being the availability of and ease of access to samples especially live human brain tissue[130]. Thus, most single cell transcriptomic studies using brain tissue have relied on either mouse brain tissue[106,109,112,131,132] or postmortem human brain tissue[90,108,110,117,118] which may be limited in their generalisability to the human brain[109,111] and accuracy[133], respectively. For example, a study examining the effect of postmortem interval on RNA integrity in rats found that mRNAs expressed at low levels were susceptible to degradation within a 24 hour postmortem interval even when stored under 4°C[133]. Thus, the use of live tissue is warranted which requires collaborations between scientists and clinicians in order for valuable surgical tissue to be harnessed where it would otherwise be discarded.

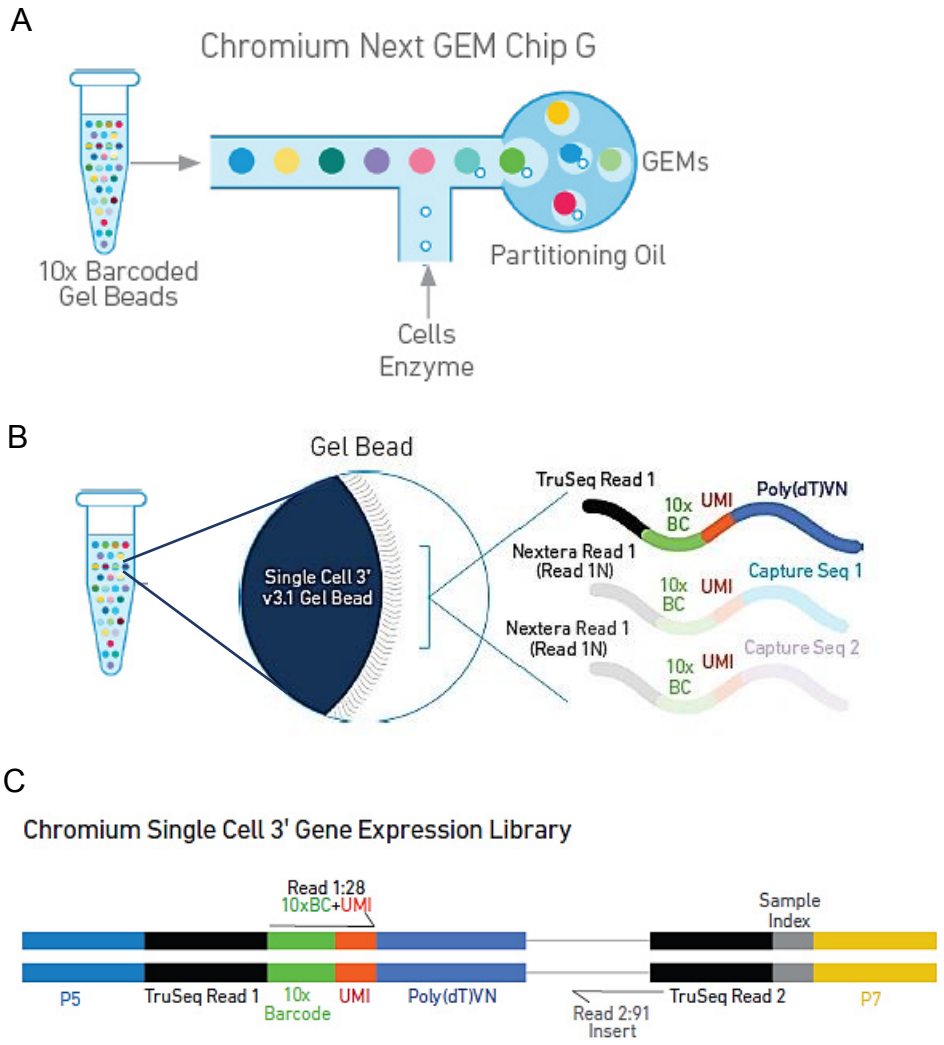


Figure 1.5. 10X Genomics single cell RNA sequencing. (A) Nuclei together with reverse transcriptase master mix (enzyme) are fed into the 10X Genomics Chromium controller. Individual nuclei are separated out and captured in partitioning oil along with a gel bed containing a unique primer set, forming a gel-bead in emulsion (GEM). (B) Each gel bead is coated with thousands of oligonucleotides containing a unique DNA barcode (10X BC) specific to the gel bead to distinguish transcripts from different cells as well as a unique molecular identifier (UMI) to distinguish each transcript from all other transcripts within a nucleus and thereby account for PCR amplification bias. (C) Schematic of the 10X Single Cell 3' gene expression library showing the P5 and P7 primers for sequencing the cDNA fragment, the TruSeq Reads 1 and 2 for Illumina sequencing, the 10X barcode (10X barcode) to distinguish transcripts from different cells, the unique molecular identifier (UMI) to account for PCR amplification bias, and a sample index to distinguish transcripts originating from different samples. Adapted from the Chromium Next GEM Single Cell 3' Reagent Kits v3.1 User Guide[134].

Using single cell technologies to explore the molecular and cellular dynamics occurring as the brain matures is likely to foster the discovery of cell type-specific diagnostic and therapeutic targets which will allow for a more directed approach to treatment[135]. Instead of drugs being designed to target the brain at large, they can be designed to target specific cell-types or subtypes within the brain resulting in more precise control of their mechanisms and fewer off-target effects or side effects. Additionally, this line of research may promote the development of treatments that are tailored to the unique requirements

of a patient by first identifying the subtle differences occurring between different groups of patients such as paediatric versus adult patients or male versus female patients. For this vision of a precision medicine to be realised, large studies involving numerous samples will be necessary in order to have sufficient power to discriminate individual variation from true biological variation between groups.

1.2. Genetic regulation of human brain maturation

Several key transcriptomic studies exist that have begun to improve our understanding of the temporal gene expression changes occurring over the course of human brain maturation[90,119–121,136]. In their pivotal and pioneering study, Kang et al. (2011) used DNA microarrays to profile the gene expression signatures of post-mortem human brain tissue from 57 donors which included a total of 1,340 tissue samples across 16 different brain regions, spanning embryonic development to late adulthood, and including multiple ethnic groups[119]. The samples were binned into various stages which included an array of distinct prenatal stages and postnatal stages. In order to investigate which factors contributed most to variation between samples, they carried out multidimensional scaling and principal component analysis which revealed that brain region and age contribute more to global transcriptome dynamics compared to the other tested variables including sex, ethnicity and inter-individual variation[119].

They went on to examine the expression trajectories of groups of genes associated with particular neurobiological categories. For example, they examined the trajectory of sets of genes associated with processes such as proliferation and migration of neural cells, development of dendrites and synapses, as well as axonal myelination. Considering that synaptic density has been shown to decrease during late childhood and adolescence[56,57], one may expect that the expression of genes implicated in synapse development would show a corresponding decrease in expression over the same period. However, contrary to expectation, the expression trajectory of synapse development genes, including *SYN1*, *SYP*, *SYPL1*, and *SYPL2* did not decline during late childhood or adolescence but instead levelled out during early childhood and remained at a consistently high level of expression[119]. One explanation for this is that genes involved in synapse development may be distinct from those involved in synaptic pruning, and thus for synaptic density to decrease during adolescence, the expression of genes involved in synaptic pruning could increase over this period without requiring a simultaneous decrease in synapse development genes. That microglia have been implicated in synaptic pruning of neuronal synapses[42,43] lends support to this hypothesis since they would likely not express genes involved in synapse development but may have a unique signature of expression responsible for their pruning activity.

Alternatively, it is possible that synapse development genes do decrease during late childhood and adolescence but that these changes are regulated post-transcriptionally or

post-translationally such that levels of synaptic mRNA or protein only differ within the actual synaptic terminals. Thus, using whole tissue samples or even individual cells may not be sensitive enough to detect these localized differences and one would have to enrich the total pool of genes for synaptosomal genes to measure this[137]. In agreement with the finding that synapse development genes reach a plateau during postnatal brain maturation, a similar DNA microarray study by Colantuoni et al. (2011) found that genes associated with synapse development, such as *GABRA1*, increased over both the prenatal and postnatal period and subsequently levelled out with no decrease in their expression observed. Nonetheless, aside from genes governing synapse numbers, the expression trajectories for genes implicated in other processes was consistent with what would be expected in terms of the observed temporal activity of the process. For example, genes associated with myelination such as *C11orf9*, *MAG*, *MBP*, and *MOG* increased dramatically over the perinatal and infancy period and continued to increase, albeit more slowly, into adulthood[119]. On the other hand cell proliferation genes, such as *HES1*, *MKI67*, and *CYCLINB2* which regulates cell cycle progression, showed the opposite trend – decreasing rapidly over the perinatal period and plateauing to a consistently low level in early childhood[119,120].

In order to survey global temporal dynamics of gene expression in the neocortex (NCX), Kang et al. (2011) determined the percentage of genes that were differentially expressed with age within broad temporal windows[119]. They found that 9.1% of all genes profiled were temporally regulated during postnatal development in the NCX from birth to adolescence while 0.7% of genes were differentially expressed across adulthood starting at 20 years of age onwards[119]. This stands in contrast to the 57.7% of genes shown to be differentially expressed across fetal development[119]. Notably, the observation that the greatest changes in gene expression occurred between the prenatal and postnatal period in terms of both the number of genes and the average effect size was a common finding across several studies[89,90,119–121,136]. This is to be expected considering that prenatal brain development is largely regulated by a genetic blueprint[62–65] whereas postnatal brain maturation is increasingly experience driven[66,67], and thus this regulatory shift would likely coincide with dramatic alterations in gene expression[119].

However, considering the myriad of micro and macroscopic changes occurring from birth to adulthood such as axonal outgrowth and myelination of axons, arborization of dendritic processes, synaptogenesis and synaptic pruning, proliferation of glial cells, alterations in metabolism, as well as the development of functionally distinct neuronal circuits[1,2,56,57,69,138,139], the abovementioned studies revealed surprisingly few and small changes in gene expression over the postnatal maturation period[89,90,119–121,136]. One possibility to explain this is that the majority of changes in gene expression occurring during postnatal brain maturation are cell type-specific and hence have been masked by an averaging out of the expression signal due to profiling populations of mixed brain cell types together. Thus, the generation of paediatric single-cell and spatial transcriptomic datasets

is necessary to address this alternative as well as the possibility that there really are few changes in gene expression dynamics over this period as the current literature seems to suggest.

In parallel to the analysis by Kang et al. (2011), Colantuoni et al. (2011) computed the rate of change of gene expression in the prefrontal cortex over a range of developmental and post-developmental stages[120]. From this, they determined that the age-dependent rates of expression change were highest during fetal development, followed by infancy and then childhood[120]. During adolescence the global age-dependent rates of expression change dropped dramatically and continued to decrease towards middle adulthood followed by a subsequent increase towards late adulthood[120]. Interestingly, with regards to inter-individual differences, global transcriptional profiles were most similar across individuals in the early years of life when age-dependent rates of expression change were the highest and these profiles subsequently diversified during the maturational years as rates of expression change declined[120]. This appears to be consistent with our understanding of postnatal maturation since environmental influences are largely responsible for shaping this period of brain maturation[66,67] with each individual exposed to a unique and diverse range of experiences that likely contribute small but specific changes to gene expression. Hence, the transcriptional signatures of each individual become gradually dissimilar with age while rates of change in gene expression decrease with age since the most drastic and universal developmental processes have already occurred.

Where Kang et al. (2011)[119] and Colantuoni et al. (2011)[120] examined overall changes in the trajectories of expressed genes across the pre and postnatal periods, Dönertaş et al. (2016)[121] developed this analysis further in their study by classifying genes into 4 categories based on how their expression changed between the paediatric (0-20 years) and adult (>20 years) brain samples. By incorporating bulk gene expression data from several studies[119–121,140], they identified a set of genes which was consistently upregulated in the developmental period and downregulated in the aging period across the various studies. Functional enrichment analysis implicated a role for these genes in certain neural functions, synaptic functions, and signalling processes[121]. However, since these studies all used whole tissue samples[119–121,140] it remains unknown whether the observed gene expression changes preferentially effect certain cell types or if all cell types are equally susceptible. Moreover, one cannot distinguish whether this reversal trend represents changes in brain cell type proportions for example due to neuronal loss, or whether it represents cell autonomous gene expression changes possibly due to the accumulation of stochastic effects. To address these limitations, the generation of cell type-specific age-series datasets spanning the lifetime is further warranted.

In an initial endeavour to fill this gap, Li et al. (2018) employed multimodal genomic methods to interrogate the transcriptomic diversity of the developing and adult brain[90].

This included generating scRNA-seq datasets using tissue samples from nine different fetal brains, generating snRNA-seq datasets using tissue samples from three adult brains, as well as performing bulk RNA sequencing using tissue from 41 individuals comprising of a range of fetal, paediatric, and adult samples[90]. Although they did not generate sc/snRNA-seq datasets for the paediatric brain, they attempted to deconvolve the bulk RNA-seq datasets into various cell types using the high resolution provided by the sc/snRNA-seq datasets[90]. Based on this they estimated changes in the relative proportions of cell types with age in the neocortex. Congruent with a subsequent independent bulk RNA-seq study by Werling et al. (2020), they determined that the numbers of neural progenitors and fetal neurons decreased over the perinatal period whereas mature adult neurons and glial cells increased over the same period[90,136]. These analyses support the notion that varying cell type proportions may contribute to the drastic changes in global transcriptional dynamics over the perinatal period. However, since these inferences were made by deconvolving bulk RNA-seq transcriptional signatures and observing changes in sets of genes associated with particular cell types[90], one cannot ascertain whether the apparent changes in cell type composition are real, or whether the changes in global transcriptional dynamics actually represent changes in the level of gene expression within a cell type (as opposed to changes in the proportion of cell types). The generation of single cell datasets across the age range under investigation, especially the generation of paediatric single cell datasets, will be integral to distinguish between these alternatives.

With a similar goal to that of Li et al. (2018), Song et al. (2021) combined 13 publicly available single cell datasets from diverse brain regions and ages and developed a "spatiotemporal cell atlas" of the human brain[141]. As there were no datasets generated using samples from individuals between the ages of 0 and 18, this age group is not present in the database. Alongside this gap, a second major limitation of this industrious study is that they did not account for batch effects such as the transcriptomic platform used to generate the datasets when examining temporal gene expression dynamics across the ages and thus one cannot be confident that the differences with age are real since the independent variable (age) is confounded by the modality used. Nevertheless, there were several interesting discoveries from this analysis including the observation that specific cell types and subtypes were enriched for genes associated with distinct neurological disorders[141]. For instance, microglia and astrocyte 3 subtypes were enriched for genes associated with multiple sclerosis whereas excitatory neuron 9 subtypes expressed genes associated with autism spectrum disorder and bipolar disorder[141]. This finding exemplifies the power of single cell analysis to provide highly resolved transcriptional signatures. If applied to pathological and control datasets, single cell technology may be capable of discerning the contribution of individual cell populations to a disease phenotype and will likely prove useful in developing more targeted therapies for treating neurological conditions. The information

from this study should of course be interpreted with caution and additional studies that use the same modality to generate the datasets is necessary to validate these findings[141].

Based on the above assessment of the current state of the literature surrounding the genetic regulation of brain maturation, it appears that the field of research will benefit immensely by the generation of single cell RNA-seq brain datasets spanning the lifetime, especially paediatric datasets between the ages of 0 and 18 years of age.

1.2.1. Long non-coding RNAs: the dark matter of the brain

Most previous transcriptomic studies examining human brain maturation have focused on the expression of protein-coding genes, with fewer studies examining the expression of non-coding genes[142,143]. As single-cell technologies target the whole genome they will facilitate the investigation of both protein-coding and non-coding genes. This is especially relevant to the study of long non-coding RNAs (lncRNAs) which have been found to show high cell type-specific gene expression patterns[144]. As the name implies, lncRNAs are a class of non-coding molecules which are greater than 200 nucleotides in length and they represent the largest class of non-coding transcripts in the genome[145]. They have been shown to carry out various regulatory roles in cells by interacting with DNA, RNA, and protein partners[146–149]. More recently, there is evidence that some lncRNAs may encode functional short peptides[150], contrary to the historical understanding that these molecules are not translated. While the functions of most lncRNAs remain poorly characterised, compelling evidence in favour of their functionality includes lncRNAs having highly conserved promoters[151,152], being dynamically regulated[124,153], localising to precise subcellular compartments[144], as well as their expression being highly region-specific[144] and tissue-specific[154] with up to 40% of lncRNAs being specifically expressed in the brain[152]. Additionally, in the past decade, numerous studies have been published characterising regulatory roles of individual lncRNAs in neural processes, for example *LOC646329* and *linc-Brn1b* in brain cell proliferation[155,156]; *RMST*, *TUNA*, and *Dali* in neuronal differentiation[149,157,158]; *Bdnf-AS* in neurite outgrowth[159]; and *Meg3* in synaptic plasticity[160]. Moreover, lncRNAs have been increasingly implicated in neurological conditions including schizophrenia, autism, epilepsy and Alzheimer's disease[147,161–163].

Based on these attributes, some researchers have hypothesised that lncRNAs RNAs may contribute to increased neuronal diversity in humans allowing increases in cognition, memory, and related abilities to emerge[164]. However, the functionality of these molecules remains a topic of controversy as many knockout and knockdown studies of individual lncRNAs have failed to show any distinct phenotypic abnormalities[165–168]. Nevertheless, this could be due to functional redundancy or that the effect size of individual lncRNAs may be relatively small[169]. Thus the analogy of lncRNAs being the dark matter of the brain holds an intriguing possibility paraphrased here from Lee et al. (2019)[170]: in the way that single elements of dark matter have incredibly small effects in the universe and yet together are

indispensable for universe integrity, individual lncRNAs may contribute non-essential, small phenotypic effects to overall brain function but in larger numbers and combinations act to provide critical regulatory roles for diverse neuronal cell types to arise thereby enabling neuronal complexity. In addition to functional redundancy and small phenotypic effects, the cell type specificity of lncRNAs may further preclude the detection of abnormal phenotypes since screening is frequently performed at a whole tissue level or even behavioural level[169]. Thus, to investigate lncRNA functions, knockout/knockdown studies may be more successful if one has an idea of which cell types are expected to be affected and screening is performed on these cells specifically. Alternatively, overexpression analysis may be warranted in lieu of knockout/knockdown studies.

Similarly to protein-coding genes, few lncRNAs have been shown to be temporally differentially expressed in the brain during postnatal maturation with one study finding as few as 8 lncRNAs being differentially expressed between paediatric and adult brains[142,143], though as many as 1500 lncRNAs showed altered expression between the prenatal and early infancy period[136,143]. It remains to be verified whether the inability to detect strong patterns of temporally differentially expressed lncRNAs in these studies is a false negative finding due to a lack of single-cell resolution and low sensitivity of the methods used to profile these genes. Evidence in support of this hypothesis comes from a study by Liu et al. (2016) who used single cell RNA-seq to profile the expression of lncRNAs in the developing prenatal human neocortex and identified 424 polyadenylated lncRNAs that were differentially expressed across developmental timepoints including *MEG3* and *DLX6-AS1* whose expression increased with developmental age[155]. Furthermore, by comparing their scRNA-seq data to equivalent bulk RNA-seq data they demonstrated that the seemingly low expression levels of lncRNAs frequently observed in heterogenous tissue could in many instances be due to cell type-specific expression of lncRNAs[155]. Developing this further, they used a large-scale CRISPRi screening method to knockdown 16,401 lncRNAs in seven diverse cell lines and found that for many of the lncRNAs, this resulted in the cell type-specific disruption of complex transcriptional pathways[171]. Together these studies illustrate the importance of investigating the expression and functions of lncRNA molecules at single cell resolution[155,171].

1.2.3. Spatial transcriptomics

In addition to single cell transcriptomics, spatial transcriptomic methods are recent developments which enable one to simultaneously examine the spatial location of thousands of genes within a single tissue section at single-cell or near single cell resolution. In 2020, spatially resolved transcriptomics was named method of the year in Nature Methods indicating the anticipated importance of this innovation for biological research[172]. Applying this method to tissue derived from the human cerebral cortex may be particularly useful considering the distinct laminar organization of the cortex whereby cells within specific cortical layers display unique gene expression patterns that appear to influence

functional properties such as neural connectivity in a layer-specific fashion[173,174]. While there are many promising spatial transcriptomic methods[175], the discussion here is centred around the 10X Genomics Visium technology since it is employed in this study.

The protocol using the 10X Genomics Visium Spatial Gene Expression kit[176] involves creating cryosections from optimal cutting temperature compound (OCT)-embedded tissue which are then placed on a specialised spatial transcriptomic slide. Each slide comprises of spots with sets of unique barcoded probes that capture mRNA. The captured mRNA is converted to barcoded cDNA, sequenced and associated back to its capture spot with bioinformatics analysis. The tissue can also be stained to reveal the underlying tissue cytoarchitecture. By combining the sequencing of barcoded cDNA and images of stained tissue sections, one can locate the region within a tissue where a specific transcript was expressed. Using this spatial data in conjunction with snRNA-seq data, one can simultaneously improve the information acquired from both snRNA-seq and Visium datasets. Visium data has comparatively lower resolution than snRNA-seq data since up to 30 cells can be captured on a single barcoded spot[177]. Nonetheless, one can deconvolve spatial transcriptomic data into various cell populations using the cell type-specific resolution provided by snRNA-seq data[178]. On the other hand, Visium data can be leveraged to annotate snRNA-seq clusters to include layer-specific information[105]. Visium data further provides information about cytoplasmic as well as dendritic and axonal gene expression whereas snRNA-seq data is limited to nuclear expression.

Maynard et al. (2021) took advantage of Visium to explore the molecular and cellular architecture of human dorsolateral prefrontal cortex (DLPFC) from three neurotypical adult samples[105]. After generating the spatial transcriptomic datasets, they devised a pseudo-bulking strategy whereby the counts for each gene were summed within each cortical layer for each sample separately to generate expression profiles defining each layer. Subsequently, using three different differential gene expression analysis methods, they showed that there were substantial differences between the various cortical layers of the DLPFC which extended beyond the expected grey and white matter differences. From this analysis, they identified novel laminar markers such as *AQP4* (L1), *HPCAL1* (L2), *FREM3* (L3), *TRABD2A* (L5) and *KRT17* (L6)[105]. They further examined whether previously published layer-specific markers such as *RELN* (L1), *WFS1* (L2), *MFEG8* (L3) and *RORB* (L4) were enriched among the Visium layer-specific DEGs. Curiously, only a subset of these published markers had a large effect size and were significantly differentially expressed in layers from the Visium datasets. This discrepancy may be due to the higher accuracy of the Visium method in quantifying transcript levels *in situ* compared to the traditional laser capture microdissection and reverse transcriptase quantitative PCR method used to quantify mRNA abundance within individual cells[179]. Alternatively, since many of the published laminar markers are derived from rodent and non-human primate studies in various brain regions and developmental time points, the discrepancy may be the result of trying to extrapolate these findings to the DLPFC of adult humans.

By overlaying both their own and publicly available snRNA-seq datasets[109,113,180,181] onto their Visium datasets, Maynard et al. (2021) showed that several neuronal subclusters within

snRNA-seq data could be defined according to specific layers in which they were expressed based on Visium data layer-specific expression signatures[105]. This included resolving excitatory and inhibitory neuronal subtypes into upper and deeper layer subclasses. Additionally, they refined the laminar annotation of several subclusters in the publicly available snRNA-seq datasets including reclassifying Ex L4 neurons in Velmeshev et al. (2019)[113] to layer 5 instead of layer 4 as well as annotating Ex4 and Ex6 neurons in Mathys et al.(2019)[180] to upper layers as opposed to deeper layers. Once again, these differences may represent region-specific, or time-point specific differences between the snRNA-seq and Visium datasets or differences in the accuracy of the methods used to annotate cell types to specific layers. This highlights the challenge ahead for researchers to develop a consensus method for annotating cell-types according to their gene expression signatures and spatial location.

Altogether, the study by Maynard et al. (2021) is an impressive feat which has provided a comprehensive framework for future spatial transcriptomic studies of the human brain such as investigating how the spatial location of gene expression as well as the laminar distribution of different cell types compares between paediatric and adult datasets.

1.3. Computational methods for processing and interpreting sc/snRNA-seq data

The establishment of cutting-edge technologies such as snRNA-seq will contribute towards tackling locally prevalent brain disorders. To this end, knowledge of the best computational tools for analysing snRNA-seq datasets is required to ensure that high-quality and meaningful information can be extracted from these datasets. Notably, as of March 2023, there are over 1400 tools for analysing scRNA-seq data, with most tools being developed for visualisation purposes[182] (Fig 1.6). An in-depth comparison of various methods is beyond the scope of this review. Nonetheless, readers are referred to several excellent reviews[127,183,184] on this topic and comprehensive evaluations of various tools[185–194]. Here, I discuss some of the main findings and conclusions from several reviews and studies with regards to current best practise for processing and analysing snRNA-seq data. I further highlight some of the main challenges the field is currently facing.

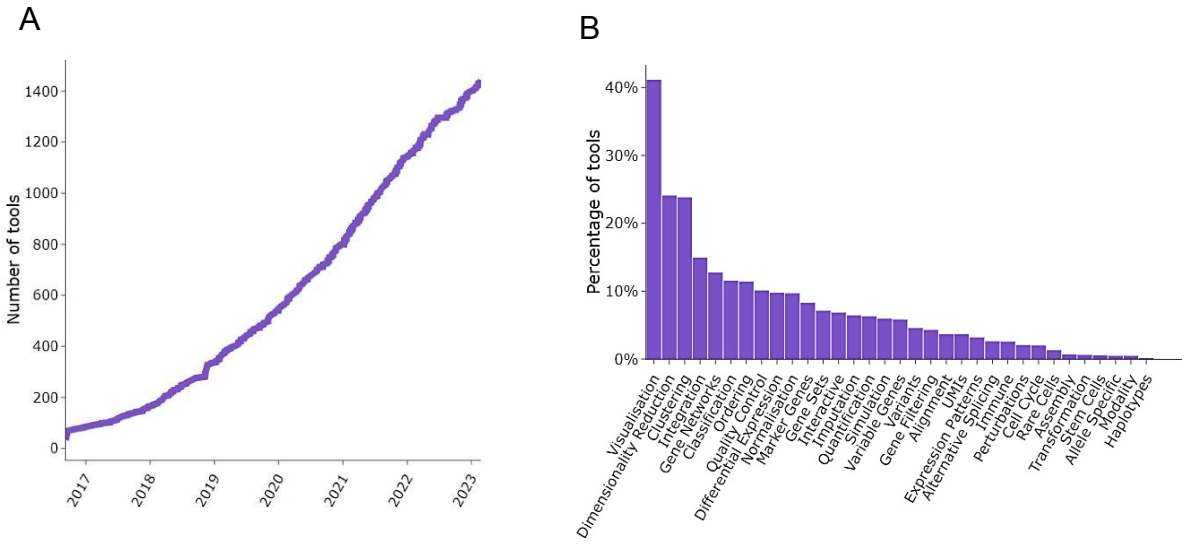


Figure 1.6. Tracking the development of snRNA-seq computational tools. (A) The number of snRNA-seq tracked between 2017 and 2023. (B) The percentage of snRNA-seq tools belonging to various categories of analysis. Figures obtained from the scRNA-tools database (https://www.scrna-tools.org/)[182].

While there are diverse strategies for processing sn/scRNA-seq datasets, most studies employ a common broad workflow[127,183] which includes: (1) demultiplexing sequencing reads and aligning them to a reference transcriptome to generate a count matrix of genes by cells (barcodes), (2) performing quality control (QC) steps to filter out poor quality cells, (3) normalizing gene counts per cell to account for differences in sequencing depth between cells, (4) correcting for batch effects and drop-out events, (5) reducing the dimensionality of the data for visualization purposes and downstream applications, (6) clustering cells according to similarities in their expression profiles, (7) annotating the resulting clusters into various cell types and/or states, and lastly (8) performing downstream analyses to make sense of the data such as cell type composition analysis, differential gene expression analysis, pseudo time trajectory analysis, and gene set enrichment analysis (Fig 1.7).

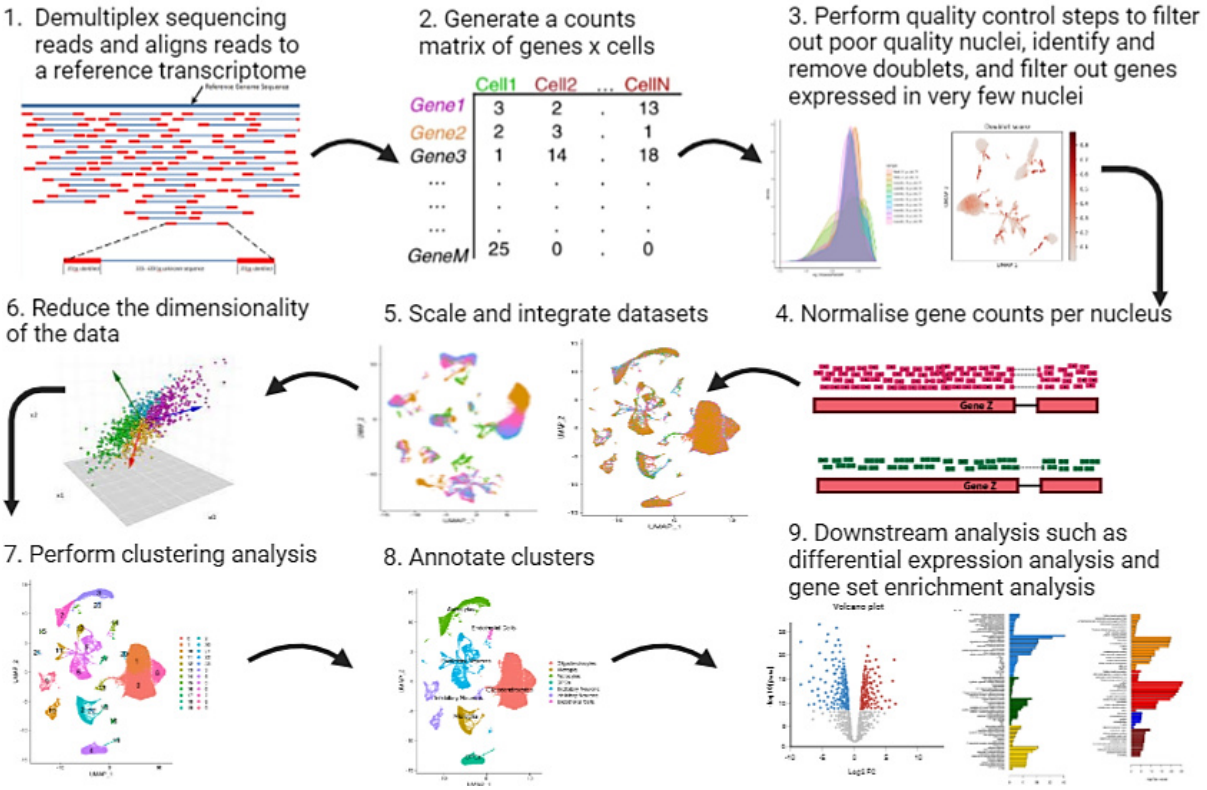


Figure 1.7. Summary of standard sc/snRNA-seq workflow. Common sc/snRNA-seq pipelines include (1) aligning sequencing reads to a reference transcriptome, (2) generating a gene by cell counts matrix, (3) performing quality control steps to obtain high-quality data, (4) normalise for sequencing depth, (5) scale and integrate the data to correct for batch effects and drop out events, (6) performing dimensionality reduction methods for visualisation purposes and downstream applications, (7) cluster the nuclei according to similarities in their expression profiles, (8) annotate the clusters into various cell types or cell states, (9) perform downstream analyses such as differential expression analysis and gene set enrichment analysis.

1.3.1. Quality control

An important consideration in the quality control pipeline is the identification of doublets. These are technical artifacts in the data that arise due to two cells being captured together and acquiring the same molecular barcode[195]. Consequently, it appears that the transcripts from the simultaneously captured cells originate from a single cell which can lead to spurious findings in downstream analysis, including the seeming existence of intermediate cell populations or cell states that do not represent real biological cell types[195]. If this occurs, it is advisable to identify these doublets and remove them from the datasets in order to draw reliable and accurate conclusions. While it is possible for more than two cells to be captured together, these forms of multiplets comprise the minority and most studies focus on removing doublets. There are two types of doublets: homotypic doublets which occur when two cells of the same type are captured together and heterotypic doublets which occur when two different cell types are captured together[195]. In theory, one can account for homotypic doublets by normalizing for sequencing depth as this will adjust the inflated gene counts to resemble that of a single cell from that population, and so identifying homotypic doublets is less essential[192]. Unlike heterotypic doublet, homotypic doublets

usually cluster with singlets and so they are believed not to interfere with classification or differential expression analysis to the same extent as heterotypic doublets[192,196].

A coarse way of identifying multiplets is to examine whether the number of transcripts associated with each molecular barcode far exceeds some measure of central tendency for the number of transcripts across all barcodes in the dataset. However, this is not always accurate since some cells or cell types may have much higher expression counts than others due to either technical or biological differences. Consequently, more advanced doublet identification tools have been designed for this purpose of accurately classifying a barcode as a doublet such as Scrublet[195], DoubletDecon[197], and DoubletFinder[196]. The principle behind each of these tools is similar in that they all attempt to simulate doublets in the data by randomly combining pairs of nuclei and then performing statistical analyses on each individual barcode to determine the probability of it being a doublet based on its similarity to simulated doublets. Where Scrublet uses the raw, unprocessed datasets[195], DoubletDecon and DoubletFinder require prior clustering of the data into various populations such that only nuclei from different clusters are paired together during doublet simulation (simulates heterotypic doublets)[196,197]. This likely contributes to the higher sensitivity of these tools compared to Scrublet. Overall, a benchmark study comparing nine tools determined DoubletFinder as the most accurate tool in terms of its impact on downstream analyses whereas DoubletDecon was found to be oversensitive in its calling of doublets leading to false positives[192]. An independent study confirmed this finding for DoubletDecon and showed that contrary to this, Scrublet had a low sensitivity and high specificity[197]. DoubletFinder represents a good balance between DoubletDecon and Scrublet with regards to the sensitivity/specificity trade-off. Nevertheless, it is recommended that multiple doublet identification tools are used in conjunction and the intersection of barcodes identified by the different tools can be used to select barcodes for removal[192].

Another important consideration during QC is the contamination of barcodes with ambient RNA. This occurs when lysed cells release their RNA into the cell suspension and this cellular debris gets captured either with unlysed cells or in its own droplet[198]. When using nuclei, the presence of mitochondrial genes is often an indication of such contamination since mitochondrial genes are not expressed in the nucleus. There are several software packages such as SoupX which have been designed with the intention of identifying and removing the contribution of cell-free mRNA to a cell's gene counts[198]. Essentially this is achieved by using empty droplets with no cell to determine the expression signature of ambient RNA, estimating the proportion of RNA in each cell that is ambient, and lastly adjusting the expression of each cell appropriately by using the two previously determined metrics.

Other QC steps include both cell-level and gene-level filtering. Cell-level filtering involves removing poor quality barcodes which have both a low number of genes being expressed

and a low total number of transcripts[199]. The ratio of genes to transcripts can also give an indication of the complexity of a cell and a threshold can be used to remove cells of low complexity if one expects to see cells of high complexity in the data as is the case with brain cells[199]. On the other hand, gene-level filtering involves removing genes that have zero counts in all cells as well genes which are only expressed in a small percentage of cells from the dataset[199].

1.3.2. Normalization and scaling

Following QC, normalization is performed to account for differences in sequencing depth between cells which can be achieved in a variety of ways such as dividing the counts for each gene in each cell by the total gene count for that cell, multiplying this by a factor of 10 000, and taking the logarithm of this value[200]. Subsequently, one can choose to centre and/or scale the gene counts by adjusting the counts such that mean expression of a gene across all cells is zero and dividing the adjusted counts by the standard deviation such that the variance across cells is 1[201]. Scaling serves as a measure to ensure that genes are weighted equally in downstream analysis such as dimensional reduction and clustering. Since highly expressed genes usually have a greater range of expression than lowly expressed genes, their standard deviations are usually also higher, and they tend to contribute more to variation in the data than lowly expressed genes. However, if one were to take the standard deviations of both highly and lowly expressed genes as a proportion of the magnitude of their expression, the values may be more comparable suggesting that variability can be a characteristic of both lowly and highly expressed genes. Thus, scaling can be thought of as equalising the playing field for both highly and lowly expressed genes to contribute to variation between cells and ultimately between samples.

According to a study by Hafemeister and Satija (2019)[202], the normalization and scaling procedure discussed above may be insufficient depending on the abundance of a gene as it assumes (incorrectly) that the same scaling factor can be applied for all genes. Specifically, they found that only low and medium abundance genes were effectively normalised and that the adjusted variance of high-abundance genes was much higher for cells that had a low total amount of RNA transcripts compared to cells with a high RNA content[202]. This indicated that genes needed to be normalized according to their abundance level. To address this, the authors devised a regularized negative binomial model (sctransform) which includes cellular sequencing depth as a covariate[202]. They demonstrated that the model can successfully eliminate the dependence between the magnitude of expression of a gene and its expression variance. As a result, variation due to technical effects is removed while true biological variation is retained such that highly variable genes can be identified irrespective of their relative abundance.

1.3.3. Correcting for technical effects, biological effects, and drop-out events

Although normalization and scaling may partially account for technical artifacts in the data, some technical effects such as batch and even a count depth effect may remain depending on the scaling method used. Additionally, biological effects which are inherent in the data such as cell-cycle differences between cells also remain and if these effects are not the variable under investigation they may contribute sources of unwanted variation to the data. One method of accounting for technical and/or biological effects is to regress out these sources of variation during the scaling step. It is recommended that if regression is performed, it is done on all covariates simultaneously so as to take into account dependencies between covariates[127]. An important consideration is that functions such as sctransform[202] are performed on single datasets individually and not across multiple datasets and so only within-sample/between-cell variation can be regressed out whereas between-sample variation cannot be. For example cell-cycle scores, cell sequencing depth, and mitochondrial-to-normal gene ratios (mitoRatios) can all be regressed out since these vary within a single dataset whereas factors such as batch, sex, or age cannot be regressed out at this point in the analysis since they are variables which differ between datasets.

Nonetheless, batch removal tools such as Combat[203] can be used for this purpose of accounting for between-sample variation. This makes use of a linear method in which batch effects are adjusted for, using both the mean and variance of the data[204]. As an alternative to batch removal tools, which usually use linear methods to correct for batch effects between multiple samples, one can also choose to perform data integration which uses non-linear methods to correct for batch effects. For example, Seurat v3 offers an integration solution which uses canonical correlation analysis to reduce the dimensionality of the data and identify correlated features between datasets[205]. A function is then performed to identify pairs of cells between datasets with similar expression profiles known as mutual nearest neighbours and these are used as “anchors” to align the datasets in reduced dimensional space by adjusting the counts data. In a benchmarking study, this Seurat integration method together with Harmony[206] and LIGER[207] were the top methods assessed for batch correction in terms of their ability to successfully mix batches and simultaneously preserve the ability to distinguish between different cell types[208]. Both linear and non-linear types of batch removal techniques are done prior to clustering and render datasets more comparable for clustering analysis and annotation. Essentially, these methods may improve the annotation of multiple datasets by ensuring that cells with similar expression profiles end up clustering together and are thus given the same annotation, regardless of which sample they come from. This in turn gives one greater confidence that downstream applications such as DE analysis between conditions are reliable since like is being compared to like (i.e microglia to microglia and astrocytes to astrocytes).

However, it is important to recognise that batch removal and data integration methods modify the count data resulting in some negative values generated which complicates DE

analysis. DE analysis tools such as DESeq2 require raw gene counts as input which are positive values and so using negative values as inputs undermines some of the assumptions of the statistic tests implemented[209]. Moreover, modifying the underlying data can obscure important biological differences between samples if the data is overcorrected. Thus, for DE analysis it is recommended that, instead of using the batch corrected expression values, one accounts for technical and biological effects by including them as covariates in the design formula, which models the effect size of the variable without altering the underlying data[127]. These covariates are then taken into account when statistical tests are performed and influence the size of determined p-values[209]. Nonetheless, batch correction can still be performed prior to DE analysis for the purpose of improved clustering and annotation (which is included as metadata) after which one can revert to the original raw counts for DE analysis.

A standout characteristic of scRNA-seq data is the abundance of zero counts in the data. This has been termed a “drop-out event” to capture the notion that the gene is expressed highly or at a moderate level in some cells but not expressed at all in others[210]. The reason for drop-out events has notoriously been attributed to technical effects such as inefficient RNA capture and differences in the platform used (droplet-based vs plate-based)[210,211]. However, there is persuasive evidence that UMI-based methods do not in fact generate higher drop-out events due to technical artifacts than would be predicted given current levels of RNA capture whereas non-UMI based methods do[210,211]. For UMI-based methods, the observed zero-inflation is instead more likely to be caused by biological heterogeneity than technical effects and thus represents true zero counts[210]. For non-UMI based methods, which do appear to have a higher rate of drop-out events than predicted, this has been attributed to PCR amplification bias which is minimised when using UMIs[210]. While capture efficiency and sequencing depth are not the main contributors of zero-inflated data, increasing these can reduce the number of zero counts in the data especially for lowly expressed genes where this problem is exacerbated[210]. This is true regardless of whether UMI or non-UMI based methods are used. Computationally, there are numerous strategies which have been developed to handle a high number of zero counts in scRNA-seq data. Many of these strategies are the same data integration methods used to correct for batch effects between multiple datasets such as Seurat integration.

1.3.4. Dimensionality reduction, clustering, and annotation

Prior to clustering analysis, dimensionality reduction is required. High dimensional data refers to data where the number of features exceeds the number of observations. Single cell data has an incredibly high number of dimensions since the number of genes being expressed and the number of cells expressing the genes far exceeds the number of samples. However, the variation between cells and samples can be sufficiently captured in far fewer dimensions than the number of expressed genes. Dimensionality reduction techniques attempt to find these dimensions and thereby describe the structure of the data in fewer

dimensions. This facilitates the summarization of the most important features that capture the variability within the data as well as the visualization of this variability in two-dimensional space. There are numerous dimensionality reduction techniques designed for this purpose, including uniform manifold approximation projection (UMAP) which is a method that essentially projects data in high dimensional space and then optimises a low dimensional graph to visualise the data in two-dimensions while preserving the structure of the high dimensional graph[212]. In a comparative study of 10 different dimensionality reduction techniques, UMAP was found to have the highest stability and was highly capable of separating cells into distinct types based on their expression profiles[191]. An alternative method, t-distributed stochastic neighbour embedding (t-SNE), had the highest performance overall and was evaluated to have the highest accuracy[191], however for visualisation purposes it may overstate differences between cell populations[127]. Thus, for the purpose of clustering and annotation, UMAP may be more suitable. Unlike UMAP and t-SNE which use non-linear methods, principal component analysis uses linear methods and can be implemented as a standard for summarization, however its accuracy decreases considerably with heterogenous data[127].

Clustering analysis is used to group cells according to similarity in their expression profiles and when visualising this process in two dimensions, it can be thought of as demarcating the boundaries between cells with similar identities. Cells are allocated to clusters based on algorithms which minimise the distances between clusters and identify densely populated regions in the reduced dimension space[213]. The resolution of clustering, which controls the number of clusters generated, can be adjusted depending on the goal of the investigation. For example, if one intends to identify cell states or novel subpopulations of cells then a higher resolution of clustering can be selected which will generate a large number of clusters in total[200,214]. Alternatively, if the purpose of the study is to see how expression profiles differ between broad populations of cells than a lower resolution of clustering can be chosen resulting in fewer clusters in total and simplifying the annotation process.

Following clustering, a marker identification step is usually performed which outputs the top marker genes defining each cluster based on differential gene expression testing between each cluster with every other cluster. The annotation of clusters into specific cell populations represents a significant challenge in the processing of snRNA-seq datasets as it largely relies on the existence of well-defined gene expression signatures of various cell types for the tissue under investigation. Fortunately, there are now cell atlases describing the diversity of brain cell type gene expression for both mouse and human[108–110,112] with recent comprehensive analysis of the human brain spanning multiple samples[116] and multiple brain regions[215].

Strategies for annotating clusters can be broadly summarised into manual and automated approaches[127,216]. Manual annotation can either involve comparing the top marker genes for each cluster to cell type-specific marker genes from the literature and cell atlas databases or vice versa whereby known marker genes are searched for in the data to see whether they are expressed in specific clusters. Automated annotation can also be split, either into automated annotation tools[217,218] or label transfer functions[205]. Both of these approaches make use of reference databases to annotate clusters, however, where the former method only queries the top marker genes for each cluster from the datasets, the latter method queries the entire expression profile of each individual cell[205], and so label transfer may be more accurate.

Ideally both automated and manual annotation tools should be used together since reference databases may not have the exact same cell types as the dataset being studied and thus one must rely on one's own discretion to determine whether a particular annotation makes sense or not. An interesting development in cluster annotation analysis is the use of machine learning methods to improve annotation. A study by Aevermann et al. (2021) describes a machine learning method to identify the minimal combination of marker genes that optimally defines clusters based on weighing up certain parameters including whether a given marker is unique to a cluster and whether it is highly expressed in all cells of that cluster[219]. This may prove useful for annotating cells by increasing the accuracy of the marker genes determined for a cluster. Moreover, it may contribute to the discovery of novel cell type-specific markers which can be added to consensus databases which formalise the knowledge of accurate markers for the classification of cell types, cell subtypes, and cell states across multiple brain regions. One exciting discovery from this study is that 24% of the determined marker genes for the human middle temporal gyrus were long non-coding RNAs suggesting that these may be contributing to cell-type-specificity[219].

1.3.5. Differential gene expression analysis

When processing single cell data, differential gene expression analysis can be performed between unannotated clusters, annotated clusters, or between groups of samples which are part of different conditions or experimental groups. The focus of the discussion here is the last option of performing DE analysis between different conditions for each annotated cell type individually. Researchers have approached this analysis using two types of strategies: either creating new DE analysis tools for specifically handling single-cell data or repurposing bulk RNA-seq DE analysis tools such that they can be applied to single-cell data (referred to as pseudobulk differential gene expression analysis)[127]. Single cell DE methods were designed to deal with specific characteristics of scRNA-seq data, such as the abundance of zeros and cellular heterogeneity, while pseudobulk methods are better suited for accounting for biological variation between samples[127].

Both approaches have significant challenges which have only begun to be addressed recently. Pseudobulk methods sum or average the gene counts across an entire cluster for each sample separately which does not take advantage of the high resolution of single-cell data as data points and information is lost in the process[220]. Essentially, the within-sample variability that exists between cells of the same sample is not accounted for. Importantly, if this within-sample variability exceeds the between-sample variability (as is typically the case with single-cell data), the aggregation or averaging of counts per sample results in inaccurate estimations of the variation within the group. Consequently, pseudobulk analysis is underpowered leading to false negative results[220]. This problem is exacerbated when the number of cells per sample differs considerably. The most effective way to increase the power is to increase the sample size (i.e have more biological replicates not more cells).

On the other hand, the main problem with the single-cell DE methods is that most of these tools have treated cells as samples when in fact they are subsamples or pseudoreplicates[186]. This violates statistical assumptions. For example, it assumes that samples are independent of each other, but they are not since cells from the same sample share a genetic background[186]. The consequence of this is that within-group variation appears to be incredibly low leading to a high false positive rate because biological variation between samples in different conditions is mistaken for a true difference between conditions[186]. Highly expressed genes are more susceptible to falsely being identified as differentially expressed using single-cell DE methods as these are generally more variable between replicates than lowly expressed genes and so small differences between groups in the expression of these genes is incorrectly attributed to the independent variable under investigation[186]. Once again, increasing the number of biological replicates helps to combat this problem whereas increasing the number of cells per replicate only worsens it[186]. The failure to properly account for dependency between cells from the same sample may explain why comparative studies of DE methods find pseudobulk methods to fare equally well or even better than single cell methods[187].

One suggestion to get around this problem is to use generalized linear mixed models, which treat each cell as a sample but include “individual” as a random effect thereby accounting for the correlation between cells sampled from the same individual[186]. Software packages such as MAST are already well-established and have shown to successfully control for type I error rates (false positives) when “individual” was adjusted for as a random effect[221]. However, the problem of a high false positive rate persists when the sample size is small[222]. To address this, Zhang et al. (2022) proposed a new method for differential expression analysis using single cell data which simultaneously accounts for the dependency between cells from the same individual as well as variation between cells of the same sample[222]. This novel tool, Individual level Differential Expression Analysis for ScRNA-seq data (IDEAS), first models the distribution of gene expression across cells from the same individual, then

compares these distributions between any two individuals, and subsequently evaluates whether within-group differences are smaller than between-group differences to decide if a gene is differentially expressed between conditions[222]. IDEAS also accounts for two types of covariates: cell-level covariates such as sequencing depth as well as sample-level covariates such as batch, sex, and age[222]. This enhances the biological signal of interest since both biological and technical variability are taken into account when making statistical inferences. However, it is important to remember that in any regression method which includes covariates, there are dependencies between biological effects and so by adjusting for one variable one may inadvertently mask the effects of another variable including the independent variable under investigation. Thus, one must ensure that the model design matrix is full rank meaning that the variable distinguishing the groups of interest is not confounded by other variables and likewise, covariates to be regressed out such as batch must not be confounded by other covariates such as sex[209]. Taken together, IDEAS represents a promising alternative to previous approaches as it exploits the information provided by single cell data without violating statistical assumptions due to improper identification of the experimental unit (i.e treating cells as samples)[222]. However, it has not yet been optimised to perform a multigroup comparison and so in instances where there are multiple levels for the variable of interest, a pairwise comparison between each level is required followed by correcting for multiple testing.

Until recently, no method had been described to specifically identify genes that change their expression in time-series scRNA-seq experiments. While time-series data has typically been analysed using pseudotime trajectory tools[223–225], these approaches are limited for identifying genes that changes as a function of time[226]. Essentially, trajectory inference methods attempt to order cells along trajectories in order to identify biologically relevant transitions between cell states or subpopulations. To do this, they use unsupervised methods which order cells according to the greatest source of variation in the data[223,224]. However, genes that are varying over the time series are not necessarily the genes that contribute the most variation between cells so these methods are not always suitable to identify time-series relevant genes[226]. To address this gap, Macnair et al. (2022) designed a tool known as Psupertime which uses a supervised regression model to identify genes that vary coherently across time-series data by providing the time-series information as input[226]. This circumvents the problem of trajectory inference methods not ordering cells in a way that recapitulates the time-series order.

To fully exploit the power of single cell analysis, post hoc computations can be performed with the significant DEGs, such as computing the percentage of nuclei expressing each gene for each condition. This will allow one to distinguish whether a gene is differentially expressed because of differences in the level of expression within individual cells or simply because more cells express the gene in one condition.

1.3.6. Gene Ontology and pathway enrichment analysis

To interpret differential expression analysis results, gene set enrichment analysis can be performed which provides insight into the putative functions of genes. This makes use of databases containing information regarding the functions of genes which have been defined according to structured, controlled vocabularies. For example, the Gene Ontology (GO) Consortium has generated a controlled vocabulary comprising of more than 30 000 precisely defined phrases referred to as GO terms which annotate genes according to their functions at a molecular level, the biological processes they are involved in, and the cellular locations where they function[227]. Gene set enrichment analysis (GSEA) determines whether a gene set from a database such as the GO database is significantly enriched in an input list of genes from omics experiments, thereby providing information of the possible functions of genes of interest[228]. There are numerous tools which have been developed to perform GSEA, including the Broad Institute's GSEA tool[229], gProfiler[230], and EnrichR[231]. Although these tools provide insight into gene functions, the definitions were established based on specific evidence from the literature usually within a particular cell line, tissue, or animal model. Thus, in order to draw any confident conclusions regarding a gene's function, further wet-lab validation of candidate genes is required.

While there are numerous computational tools which have not been described here, I have provided an overview of a typical sc/snRNA-seq workflow, highlighting some of the key considerations and popular tools for carrying out this analysis.

1.4. Relevance of the research in the South African context

Unlike neuropsychiatric illnesses which often have a strong heritable component, there are many neurological disorders which are sadly preventable but remain prevalent in developing countries such as South Africa[232]. Poverty is a leading risk factor for preventable neurological disease with violence, poor medical services, and environmental factors such as lack of hygiene, epidemics, trauma, and natural disasters being associated risk factors[232,233]. In South Africa, neurological complications as a result of infections such as tuberculosis, HIV, and streptococcus are the dominant neurology cases[234–236]. In terms of the paediatric population specifically, seizures as a complication of neuroinfection, traumatic brain injury[95,237,238], and foetal alcohol syndrome remain prevalent, with the Western Cape province of South Africa having the leading rate of foetal alcohol syndrome globally[239]. Devastatingly, South Africa also faces a high rate of cerebral palsy incidence compared to the global average[240,241] – a fact which has been attributed to medical malpractice during childbirth[242].

While there are major efforts to treat and combat these acquired neurological conditions, there are few studies examining these conditions at the level of transcriptome-wide gene expression in the paediatric brain, with even less known about cell type-specific dysregulation of gene expression in both the maturing and adult brain[243–247]. Nonetheless,

there have been several insights into the aetiology of specific conditions at the level of gene expression. For example, a scRNA-seq study in a mouse model of acute concussive brain injury implicated specific genes and cell types in the pathogenesis of traumatic brain injury (TBI) which included the dysregulation of metabolic genes in astrocytes and neurons whereas amyloid genes were dysregulated in ependymal and endothelial cells[122]. Notably, altered metabolism has also recently been implicated in microglia of paediatric TBI patients[248] which is corroborated by an independent single-cell study in a mouse model of mild TBI[249]. This line of research may promote the generation of novel cell type-specific therapies such as drug-conjugated dendrimer nanoparticles[250] to manipulate microglial function which may be useful in treating the numerous disorders where microglia have been implicated such as HIV-associated neurocognitive disorder[251,252] and meningitis[253].

Future endeavours by South African paediatric researchers can focus efforts into understanding the molecular and cellular basis for various neurological disease phenotypes relevant to the population using single-cell methods. Considering that by 2050 37% of the world's children will grow up in Africa[254], studying the African paediatric population and understanding how neurological disorders manifest differently in this population compared to adults has never been more pertinent.

1.5. Research aims and objectives

1.5.1. Research aims

The main aim of this study is to investigate the cellular and molecular changes occurring during postnatal human brain maturation at single cell resolution with a particular focus on changes to the expression of lncRNAs over this period. Transcriptomic datasets were generated for both paediatric and adult datasets in order to identify changes occurring as the brain matures. The adult datasets served as a reference for exploring the paediatric datasets.

1.5.2. Research objectives

1. Use the 10X Genomics Chromium controller to generate snRNA-seq datasets from bio-banked ante-mortem brain tissue samples.

2. Apply bioinformatics methods to process the snRNA-seq datasets from Objective 1 alongside publicly available datasets. This includes i) filtering the datasets to retain high quality nuclei, ii) integrating and clustering the datasets, iii) annotating the resulting clusters, iv) identifying the minimal marker genes required to define and distinguish cell types, v) performing analyses to identify both coding and non-coding genes whose expression changes with age, and iv) determining putative functions of a subset of relevant DEGs using GSEA

3. Use the 10x Genomics Visium spatial transcriptomic solution to validate the expression of a subset of relevant genes which represent cell type-specific markers.

4. Perform *in silco* functional characterization of lncRNAs of interest.

1.6. Research hypotheses

While the nature of this study is largely exploratory as opposed to hypothesis-driven, I developed several loose hypotheses based on the state of the literature. I hypothesise that we will identify a greater number of DEGs with age compared to previous transcriptomic studies of brain maturation. I expect to see this increase because the analyses may expose cell type-specific expression differences between ages that would previously have been masked by an averaging out of the expression signal when measuring gene expression at the tissue rather than cellular level. Additionally, I speculate that some of the marker genes characterising cell types may be specific to either paediatric or adult samples. Lastly, I hypothesise that some of the genes whose expression changes with age may be associated with early-onset neurological conditions as well as biological processes that underlie brain maturation.

Chapter 2: Research Methodology

2.1. Live human brain tissue samples

Ethics was granted for the use of paediatric and adult human brain tissue by the University of Cape Town Human Research Ethics Committee (UCT HREC REF 016/2018; sub-study 147/2022). The human brain tissue samples used in this study were obtained during elective surgeries performed at the Red Cross War Memorial Children's Hospital and Constantiaberg Mediclinic in Cape Town, South Africa. The samples were all of temporal cortex origin and included a total of 23 samples from 12 donors (Table 2.1). Tissue was transported in carbogenated choline (1X Choline, 0.03 M NaHCO [Sigma-Aldrich, US], 0.0 1M D-Glucose [Sigma-Aldrich, US], 0.005% PenStrep [Sigma-Aldrich, US]) immediately from the hospital to the laboratory (~30 mins). Individual dissected pieces, which contained all layers of the temporal cortex from the outer layer 1 to the inner white matter, were either flash frozen in liquid nitrogen or embedded in OCT and stored at -80 °C until needed. The OCT-embedded pieces were flash frozen in a 10×10 mm^2 cryomold [Sigma-Aldrich, US] 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 datasets from Thrupp et al. (2020)[255] were downloaded from the Sequence Read Archive (SRA) database and were based on samples obtained during elective surgeries performed at UZ Leuven in Belgium.

Table 2.1. Summary of 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)[255]. The P0013 and P0015 datasets were generated by Stephanie Fillmore for her Master's dissertation.

Sample ID	Age (years)	Sex	Diagnosis	Analysis	Number of technical replicates	Technical replicate	Batch	10X Genomics Single Cell Chemistry Platform
P0001	4	M	Left temporal lobe dysplasia and epilepsy	snRNA-seq, Visium	1	T1	E	V3.1
P0018	5	F	Right anterior temporal lobectomy and removal of choroidal cyst	snRNA-seq	2	T1	D	V3.1
						T2	D	
Nuc-RM77	7	M	Epilepsy caused by brain tumor	snRNA-seq	2	T1	A	V2.0
						T2	A	
P0011	9	F	Right sided Rasmussen's with refractory epilepsy	snRNA-seq	2	T1	D	V3.1
						T2	D	
P0013	15	F	Left temporal lesion (ganglioma) and epilepsy	snRNA-seq, Visium	6	T1	B	V3.1
						T2	B	
						T3	B	
						T4	B	
						T5	C	
						T6	C	
P0029	15	M	Left temporal lobe epilepsy	snRNA-seq, Visium	1	T1	G	V3.1
Nuc-RM102	20	F	Therapy resistant epilepsy	snRNA-seq	2	T1	F	V2.0
						T2	F	
Nuc-RM95	24	F	Epilepsy caused by brain tumor	snRNA-seq	2	T1	F	V2.0
						T2	F	
P0028	26	F	Left medial temporal lobe/ hippocampus glioblastoma	snRNA-seq	1	T1	E	V3.1
P0015	31	M	Temporal lobe epilepsy	snRNA-seq, Visium	2	T1	C	V3.1
						T2	C	
P0026	41	F	Right temporal neocortical epilepsy,	snRNA-seq	1	T1	G	V3.1
Nuc-RM101	50	F	Therapy resistant epilepsy	snRNA-seq	1	T1	A	V2.0

2.2. Nuclei isolation for single nucleus RNA sequencing

Nuclei were isolated according to a protocol adapted from Habib et al. (2017)[256] and the 10X Genomics nuclei isolation protocol (CG000124, User Guide Rev E)[257] (Fig 2.1). Frozen brain tissue was homogenised in a dounce-homogeniser containing 2 ml ice-cold lysis solution (Nuclei EZ Lysis Buffer [Sigma-Aldrich, NUC101] on its own 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 mins on ice.

The sample was centrifuged at 500 x g for 5 mins at 4 °C after which the supernatant was discarded and the nuclei resuspended in 3 ml ice cold nuclei suspension buffer (1xphosphate-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 mins 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 mins at 4 °C after which it was mixed with 3 ml blocking buffer and centrifuged at 300 x g for 5 mins at 4 °C. The supernatant was removed and the nuclei were resuspended in 2 ml clean blocking buffer. The sample was transferred to a 2ml microtube and placed on a Dynamag magnet for 15 mins at 4 °C. The supernatant was transferred to a new microtube and stored on ice. Equal volumes of trypan blue and nuclei suspension were mixed and loaded onto a haemocytometer and a brightfield microscope was used to count nuclei. The concentration of nuclei was subsequently adjusted by either concentrating or diluting in an appropriate volume of blocking buffer to obtain a concentration of ~1000 nuclei/µl.

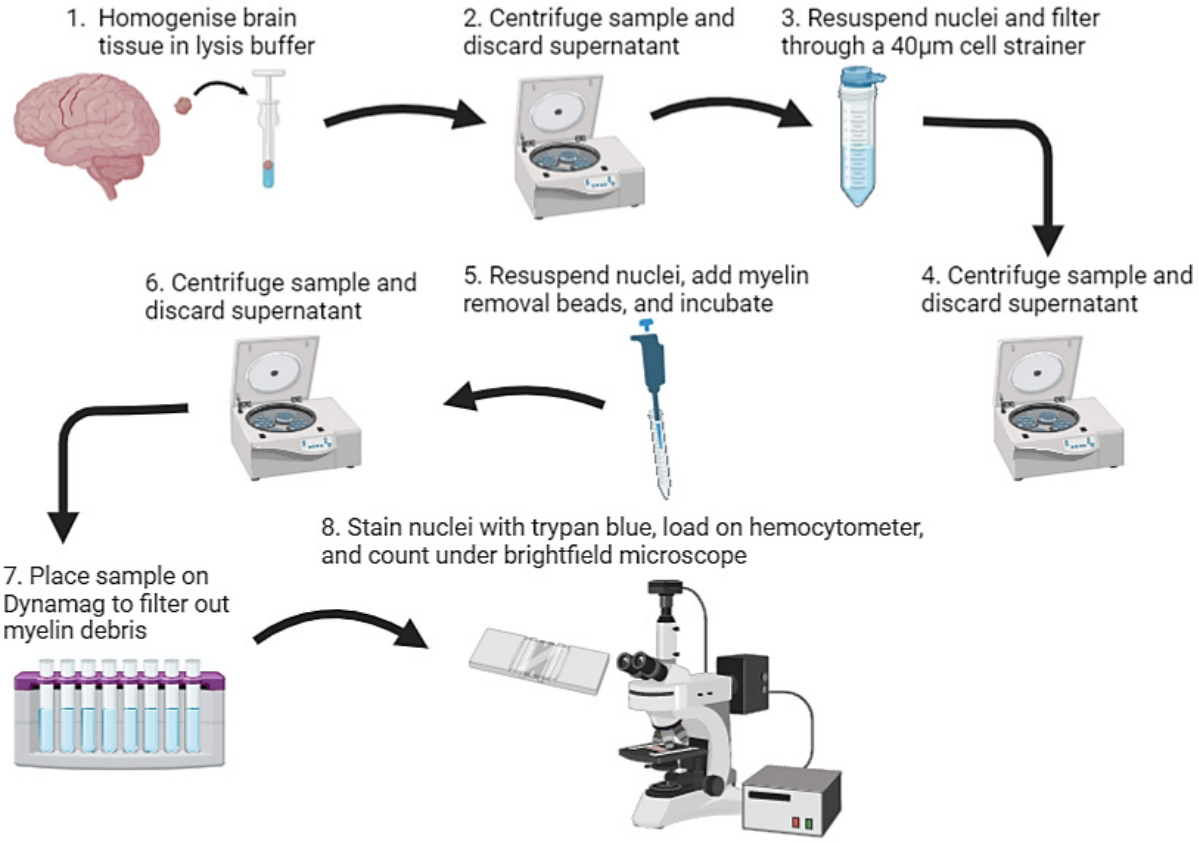


Figure 2.1. Schematic of the sample preparation workflow for snRNA-seq platform. (1) Flash frozen brain tissue piece was homogenised in lysis buffer to rupture cell membranes and release the nuclei from cells. (2) The sample was centrifuged to pull down nuclei and the supernatant containing cell debris discarded. (3) Resuspended nuclei were filtered through a 40 µm cell strainer to further remove debris. (4-6) Two centrifugation steps were performed and myelin removal beads were added to the nuclei suspension. (7) The Dynamag magnet was used to bind myelin debris attached to the myelin removal beads and the supernatant obtained. (8) Nuclei were stained with trypan blue and loaded onto a hemocytometer. A brightfield microscope was used to visualise the nuclei and manual counting was performed to estimate nuclei concentrations. *Figure compiled with Biorender.*

2.3. 10X Genomics snRNA-seq library preparation

snRNA-seq library preparation (see Fig 1.5) was carried out using the 10x Genomics Chromium Next Gen Single Cell 3' Reagent Kit (v3.1)[134] according to manufacturer's protocols (CG000204, User Guide Rev D). The 10X Genomics Chromium Controller was loaded with partitioning oil, gel-beads, nuclei, and master mix (reverse transcription [RT] reagent, template switch oligo, Reducing Agent B, and RT Enzyme C). The volumes of nuclei suspensions loaded were determined using the Cell Suspension Volume Calculator Table in the 10x Genomics user guide to target 10 000 nuclei per sample. Individual nuclei are captured in an oil droplet together with a gel-bead to form gel bead-in-emulsion (GEMs). Each gel bead comprises of thousands of oligonucleotide probes which have a poly(dT) primer sequence for capturing transcripts by their polyA tails, a UMI, and a 10X barcode specific to the gel bead (Fig 1.5). The transcripts in each nucleus are converted to barcoded, full-length cDNA during the GEM-RT incubation step. Following GEM incubation, the cDNA was amplified using 11 PCR (polymerase chain reaction) cycles yielding a sufficient mass for library construction. To assess the quality of the libraries, their fragment size distributions were determined using the Agilent TapeStation or Agilent Bioanalyser at the Central

Analytical Facility (CAF, University of Stellenbosch) (Supp Fig 2.1-2.3). Qubit analysis was also performed to quantify cDNA concentrations (Supp Table 2.1).

Subsequently, an enzymatic fragmentation step and a size selection step was performed using Solid Phase Reversible Immobilization Methodology (SPRI) select reagent (Beckman Coulter, US). This was done to select cDNA amplicons that are of optimal size for Illumina sequencing. A sample index and sequencing primers were added to each of the libraries via fragmentation, end repair, A-tailing and Sample index PCR. Based on the cDNA yield determined by the Qubit analysis (Supp Table 2.1), 13 PCR cycles were selected to amplify the libraries, after which quality control was performed (Supp Fig 2.1-2.3) as was done prior to the fragmentation step. The final libraries comprised of P5 and P7 primers for sequencing the cDNA fragment, TruSeq reads 1 and 2 for paired-end Illumina sequencing, a 16 base pair (bp) 10X barcode distinguishing transcripts from different nuclei, a 12 bp UMI distinguishing each original transcript for accurate transcript quantification post-sequencing, and the sample-index to enable sample multiplexing during Illumina sequencing (Fig 1.5). cDNA libraries were sequenced by Novogene (Singapore) on either the Illumina HiSeq or NovaSeq system using the Illumina High Output kits (150 cycles).

2.4. snRNA-seq bioinformatics analysis

The snRNA-seq datasets were processed using a pipeline adapted from the Harvard Chan Bioinformatics Core[199]. This included data pre-processing, quality control steps, data normalisation, scaling, integration, clustering analysis, marker gene identification, and differential expression analysis (see Fig 1.7).

2.4.1. Read alignment and gene expression quantification

Fastq file reads obtained from Novogene were aligned to a 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, https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) (Code availability, script 1). The inclusion of introns was specified in the count function. This was done for all datasets included in this study (those generated in our lab as well as the publicly available datasets). An automatic filtering process was performed to remove barcodes corresponding to background noise which have very low UMI counts and likely represent GEMs which captured ambient RNA from dead or lysed cells. The quality control outputs from this analysis are in Supp Table 2.2 and Extended Data 1.

2.4.2. Quality control

The filtered gene barcode matrix for each sample was imported into R using the Read10X function from the Seurat (v.2.0) package[200]. Nuclei-level filtering was performed to remove poor quality nuclei according to their number of unique molecular identifiers (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[199]. 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. Additionally, three doublet removal tools namely DoubletFinder[196] (Code availability, script 3), DoubletDecon[197] (Code availability, script 4), and Scrublet[195] (Code availability, script 5) were used to identify doublets for each dataset individually. Doublets are technical artefacts due to two nuclei being captured together during GEM generation and being labelled with the same the same molecular barcode which can confound downstream analysis[195]. The sample-specific parameters of each of the tools were adjusted according to the guidelines. For Scrublet, the threshold for doublet removal was determined by the number of doublets corresponding to the minimum between the two modes of the distribution of simulated doublets. For DoubletFinder, a parameter known as the pK value, which is a measure of neighbourhood size in gene expression space, was adjusted. A neighbourhood that is too large or too small can negatively affect the ability to distinguish singlets from doublets. For DoubletDecon, the rho prime value was optimised for each dataset to control the merging of similar clusters. In general, larger rho prime values were selected in favour of separating clusters instead of merging them in order to promote accurate simulation of heterotypic doublets. 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[197].

2.4.3. Data normalization, integration and clustering

Prior to integration, principal component analysis (PCA) was performed to evaluate known sources of within-sample variation between nuclei, namely the mitoRatio and cell cycle phase (Code availability, script 6). 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 since it was a source of unwanted variation between nuclei. 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 to align similar cell types across the datasets (Code availability, script 6). The features selected for integration were genes that were shared between samples of the 3000 most variable genes identified for each sample by SCTransform. To cluster the datasets following integration, dimensionality reduction was first performed using UMAP embedding, specifying 40 dimensions (Code availability, script 7). The Seurat FindClusters function, which uses a shared nearest neighbour (SNN) clustering algorithm, was then applied at a resolution of 0.8 – producing 40 clusters.

2.4.4. Cluster annotation

Clusters were annotated at both the broad cluster level as well as the subcluster level. The broad cluster annotations were determined using a consensus annotation from several methods. Firstly, the FindAllMarkers function was applied to the integrated datasets using the default Wilcox test, a log fold change threshold of 0.25, and specifying positive markers only (Code availability, script 8). The output of FindAllMarkers (Supp Table 2.3) was used as input into an automated annotation tool, SCSA[218] (Code availability, script 9), which attempts to annotate clusters based on matching the marker genes identified for each cluster to lists of known cell type markers from the literature. A parallel automated annotation tool, scCATCH[217] (Code availability, script 10), was also applied which is based on the same principle as SCSA. In addition to automated annotation, violin plots were generated for several cell type-specific marker genes from Hodge et al. (2019)[109], Bakken et al. (2018)[129], and Lake et al. (2018)[110] to examine which clusters expressed the markers most highly (Code availability, script 11). Lastly, an independent label transfer operation was performed using Seurat's TransferData function wherein the Allen Institute for Brain Science's Smart-seq human middle temporal gyrus (MTG) dataset[109] served as a reference, while the integrated dataset served as the query (Code availability, scripts 12-13). 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. The nuclei from the reference MTG dataset were annotated into 1 of 75 transcriptionally distinct cell types that were labelled according to the broad cell type they belonged to, the layer(s) from which they were dissected, the expression of a marker gene for the broad subtype, and the expression of a marker gene for the specific subtype[109]. For example, nuclei belonging to Exc L2 LAMP LTK population were assigned to the excitatory class (Exc) based on higher expression of the excitatory marker, *SLC17A7*, compared to other broad cell type-specific markers. The nuclei were dissected from layer 2 and expressed *LAMP5* at a higher level than other excitatory subtype markers such as *RORB* and *THEMIS*. Lastly, the marker gene, *LTK*, showed the greatest difference in its level of expression in the Exc L2 LAMP LTK cluster compared to all other clusters.

By combining the various manual and automated annotation methods, the 40 initial clusters were collapsed into 7 broad cell types, namely inhibitory neurons, excitatory neurons, oligodendrocyte precursor cells (OPCs), astrocytes, oligodendrocytes, endothelial cells, and microglia (Code availability, script 14). The subcluster annotations were derived solely from the label transfer operations using the Allen MTG as the reference dataset. These included high resolution annotations resulting in 54 different cell-types (Allen_high_resolution_cluster_label). To enhance our certainty of the high resolution cluster annotations assigned, the expression of top marker genes from Hodge et al. (2019) were plotted to confirm their expression in the expected cell type. Cell composition plots

showing the relative abundance of the 54 cell types were generated comparing either the 12 samples, the 23 technical replicates, or the adult samples to the paediatric samples.

As an additional validation, similarity scores were computed to compare the transcriptomic similarity of each of the 54 query cell types to the 75 reference middle temporal gyrus cell types (Code availability, script 15). This was achieved by first sub-setting genes in our data with a beta score > 0.5. The beta score is a measure of a gene's ability to distinguish different cell types. While a high beta score is ideal to maximise the classification power of the subsetted genes, choosing a beta score that is too high would result in an insufficient number of genes being subsetted to be able to distinguish all cell types, hence a threshold of 0.5 was used. A Pearson correlation was then performed to correlate the median expression of the subsetted genes between our clusters and the Hodge et al. (2019) clusters using the gene counts from the RNA assay counts slot.

2.4.5. Cell-type marker identification

The NS-Forest tool (v3.0)[219] was used to identify small combinations of marker genes uniquely defining each high resolution subcluster for each sample individually (Code availability, script 16-18). This was performed to explore the diversity between cell sub-types and between samples without having to examine hundreds of marker genes. A random-forest model was used to select a maximum of 100 marker genes per cell type and sample based on them being both highly expressed as well as uniquely expressed within a cell type compared to other cell types. The list of markers was ranked accordingly and the top 6 genes were selected for further testing to identify the smallest combination of marker genes necessary to define a cell type and distinguish it from other cell types. This process is achieved in six main steps outlined in Aevermann et al. (2021) namely: (1) A cell-by-gene matrix comprising of annotated cell types serves as input for producing binary classification models for each cell type, (2) marker genes were identified from the models and ranked according to Gini Index (measure of the probability that a randomly chosen gene is incorrectly classified), (3) negative marker genes were removed, (4) the remaining genes were ranked by their binary expression scores and the top 6 genes were extracted, (5) decision tree analysis was used to obtain expression level cutoffs for the marker genes, and (6) F-beta scoring was used to assess the classification power of all permutations of the selected markers and the top permutation selected as the best combination of markers for defining the cell type. The number of trees chosen for this model was 50 000, 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 NS-Forest outputs to be directly compared to those from Hodge et al. (2019)[109] and Aevermann et al. (2021)[219]. To assess the relevance of these markers in terms of their capacity to distinguish different cell types, the SCT and integration methods were repeated using either a random set of genes or the NS-Forest markers as anchors for integration based on the quality control method used by Aevermann et al.

(2021)[219] (Code availability, script 19). The resulting UMAPs were compared to the original UMAP (Section 2.3.4) which used the shared highly most variable genes across datasets as anchors.

2.4.6. DESeq2 age-dependent differential gene expression analysis

DESeq2[209] was used to identify genes that were differentially expressed with age (Code availability, script 20). Samples were first binned into five different epochs as follows: epoch 1 = 4-year-old and 5-year-old; epoch 2 = 7-year-old, 9-year-old; epoch 3 = 15-year-old B1 and 15-year-old B2; epoch 4 = 20-year-old, 24-year-old, 26-year-old, and 31-year-old; and epoch 5 = 41-year-old and 50-year-old. These bins were chosen in agreement with the periods of human development and adulthood defined in Kang et al. (2011). The 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. Principal component analysis was performed on each cluster 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 + epoch was specified to treat epoch (age) as the variable of interest while the effect of single_cell_chemistry (version2 vs version3 chemistry) was regressed out. The full model thus included both epoch and single_cell_chemistry whereas epoch was removed from the reduced model. The DESeq2 function was performed using the likelihood ratio test (LRT) with the null model equal to ~single_cell_chemistry indicating that the effect of this covariate should be modelled. DESeq2 models the counts for each gene using a negative binomial distribution as described by the generalised linear equation below:

$$K_{ij} \sim NB\ (\mu_{ij}, \alpha_i)$$

where K_{ij} represents the raw count for gene i, sample j that are modelled using the negative binomial distribution, NB. The parameters μ_{ij} and α_i represent the fitted mean and a gene-specific dispersion parameter, respectively.

The raw gene counts were normalised using the median of ratios method which makes use of sample-specific size factors to account for differences in sequencing depth between samples, including differences in the number of nuclei per sample[209]. A regularized log transformation of the normalised counts was performed which moderates the variance across the mean and gene-wise dispersion estimates are then computed for each gene.

The LRT method allows for a multigroup comparison by conducting a hypothesis test comparing the fit of the full model to the fit of the null model for each gene. The only difference between the full model and the null model is the inclusion of epoch as a covariate into the full model, and thus by comparing the fit of these two models to each other, one can test whether there is a significant amount of variation that is explained by age. Briefly

this is done by computing a statistic known as the likelihood ratio (LR) which can be described as follows:

$$LR = -2\ln\left(\frac{L(m1)}{L(m2)}\right)$$

where L(m1) represents maximising a likelihood function such that under the null model, the observed expression count for a gene is most probable

and L(m2) represents maximising a likelihood function such that under the full model, the observed expression count for a gene is most probable.

DESeq2 implements an Analysis of Deviance (ANODEV) method to test whether this ratio is significantly different from zero and calculates an associated p-value based on LR following a chi-squared distribution. If the null hypothesis is rejected for a gene, this can be understood as the gene being differentially expressed across the different levels (epochs), that is to say, the level of gene expression is significantly different between at least two epochs. The associated p-values for each gene were adjusted for multiple testing using an alternative of the Benjamini-Hochberg method. A precise interpretation of adjusted p-values that are significant is as follows: the addition of epoch as a covariate in the full and not in the null model increased the log likelihood more than would be expected if the true value for the coefficient associated with epoch was zero.

The log2 Fold Changes for each pairwise comparison were obtained using the DESeq2 results function and specifying each contrast (e.g epoch1 vs epoch2; epoch1 vs epoch3 etc). Additionally, the set of significant DEGs for each cluster were grouped according to similar temporal patterns of expression using the degPatterns function from the DEGreport package[258].

2.4.7. Psupertime time-series single cell differential gene expression analysis

Psupertime[226] was used to identify a set of interpretable genes whose expression varies with age within each cell type (Code availability, script 21). To achieve this, Psupertime first performs filtering to remove genes expressed in fewer than 10% of nuclei in the cell population of interest, data denoising, correcting for drop out events, and scaling of gene expression counts. This is done for each cell type on its own where previously filtering was performed on the entire dataset for each sample (Section 2.3.2). Subsequently, ordinal regression is performed to determine a coefficient for each gene with the likelihood for the regression defined by multiple simultaneous logistic regression equations that try to separate the nuclei according to the age of the donor from which they originate *i.e if the number of donors in the time-series is K, each equation seeks to separate donor 1…k from donor k+ 1…K*. Thus, the resulting coefficients are a measure of each gene's ability to explain the order of the time-series labels. Genes with a non-zero coefficient are considered to be relevant and should vary in their expression with age. To visualise the change in gene

expression across the time-series, Psupertime computes a pseudotime value for each gene which approximately recapitulates the order of the time-series labels by multiplying the expression value of each gene by its estimated coefficient. For each nucleus, the expression value of a gene of interest can then be plotted against the pseudotime value for the nucleus. This is done for all nuclei with a cell type to visualise the change in expression with donor age.

Psupertime was applied to each cell type individually resulting in a list of relevant genes (for which the coefficient $\beta i > 0$). Notably, cell populations which had an insufficient number of nuclei in total or had an insufficient number of nuclei for at least one sample were excluded from the analysis as a generalized linear model could not be fitted to the data (the penalized factor exceeded the number of data points). For the successful cell populations, the relevant genes were then grouped into patterns of similar expression using a standard hierarchical clustering method (R's hclust method was applied to the scaled log transformed average expression values across the samples). The number of clusters chosen was k=20 unless there were fewer than 20 genes in which case the number of genes was used. Clusters with seemingly similar expression patterns were grouped together on a posthoc, observational basis.

2.4.8. IDEAS pairwise adult versus paediatric differential gene expression

IDEAS[222] was optimised to conduct a pairwise comparison between the six paediatric and six adult samples (Code availability, script 22-23). Briefly, for each cell type individually, genes expressed in < 10% of nuclei in the population were removed. The distribution of each gene's expression across nuclei for each individual was estimated using a parametric negative binomial distribution which takes into account cell-level covariates such as the sequencing depth of each nucleus. For each gene, the distance between the distributions of any two individuals was then calculated using the Wasserstein distance. A hypothesis test was then conducted to determine whether the within-group distances are smaller than between-group distances. The null hypothesis is that the expression of a given gene is not associated with the variable of interest (age group), accounting for other individual-level covariates such as batch or sex (in this case we accounted for the single-cell chemistry platform used). The output of the test is a permutation p-value for each gene (computed using a kernel regression method) which is as a measure of whether the expression of the gene is significantly associated with age group. The resulting significant genes were clustered into two groups using the hclust hierarchical clustering method (k=2).

2.4.9. Enrichment analysis

GSEA was performed on the set of DEGs identified by DESeq2 (Section 2.3.7), Psupertime (Section 2.3.8), and IDEAS (Section 2.3.9) to determine putative functions of the genes (Code availability, script 24). The analysis was performed for each cluster individually on sets of genes which were previously grouped into similar patterns of expression with age. This was

achieved using EnrichR[231] which is a database comprising of multiple curated gene-set libraries comprising of various biological terms and their associated genes. These gene-sets have been grouped into various categories such as Transcription, Ontologies, Pathways, Diseases. A gene list of interest (usually coexpressed genes as these often share a biological function) is used as input and a Fischer's exact test is performed to determine whether the input set of genes significantly overlaps with any of the gene sets in the database. The Benjamini-Hochberg method is used to adjust the computed p-values for multiple testing. For this analysis, gene lists were queried for overlap with three gene set libraries in the EnrichR database, namely GO Biological Process 2021 (genes associated with known biological processes), DisGeNET (collections of genes associated with human diseases), and GTEx Aging Signatures 2021 (genes up or downregulated with age in various tissue from the Genotype-Tissue Expression V8 dataset). To visualise the results, the output terms were first ordered by their p-values after which the top 10 terms per database were sub-setted.

2.4.10. Proportion analysis comparing paediatric to adult datasets

To exploit the information available from the single cell data, a hypothesis test was performed to determine whether there was a significant difference in the proportion of nuclei expressing the genes between paediatric and adult datasets (Code availability, script 25). The proportion of nuclei instead of the number of nuclei was taken to account for variability in the total number of nuclei per cluster per sample. The hypothesis test was performed on each cell type individually and the gene list was first filtered to remove genes expressed in fewer than 10% of nuclei in the cell type of interest across all datasets. Subsequently, either a parametric unpaired two-samples t-test or parametric unpaired two-samples Wilcoxon test was performed depending on whether the data was normally distributed. The p-values were adjusted for multiple testing using the false discovery rate method.

As an alternative method to account for possible differences in the total number of nuclei per sample, the number of nuclei were randomly downsampled to that of the sample with the fewest nuclei (Code availability, script 26). This was only performed on clusters with at least 50 nuclei per sample. The above pipeline was repeated, comparing the number of nuclei expressing the gene instead of the proportion.

2.4.11. Long noncoding RNA analysis

To further explore selected lncRNAs of interest, a 'Guilt by association' method was used to correlate the expression of the lncRNAs with protein coding genes and identify sets of positively and negatively correlated genes (Code availability, script 27). The principle behind this assumes that coexpressed genes function in the same biological pathways and by taking advantage of the known function of protein-coding genes, the function of the lncRNAs can be inferred[259]. The cell type of interest was sub-setted from the Seurat object containing a gene-by-cell matrix for the 23 merged datasets. The RNA assay and data slot

were then extracted after which the BaCo function (Bayesian Correlation) from Sanchez-Taltavull et al. (2020)[260] was applied. The Bayesian method was chosen as it has been shown to be a robust gene-by-gene similarity measure for single cell data[260]. The coexpressed protein-coding genes corresponding to the lncRNA of interest were extracted and ranked according to their correlation score. The top 50 positively and negatively associated genes were sub-setted and GSEA was performed using EnrichR to determine biological processes that these genes are involved in.

The interaction partners of lncRNAs can also inform their putative functions. Here, the computational tool Fasim-LongTarget (https://lncRNA.smu.edu.cn)[261,262] was used to predict possible DNA interacting partners of the selected lncRNAs. A lncRNA can form a triplex molecule by binding to a duplex DNA sequence through Hoogsteen or reverse Hoogsteen base pairing[263]. LncRNAs can bind at sites proximal or distal to a gene[264] and can regulate gene expression in several ways including recruiting epigenomic modification enzymes to DNA binding sites[265]. Fasim-LongTarget predicts putative DNA binding motifs of a lncRNA by querying its sequence against the region 3500 bp upstream and 1500 bp downstream the transcription start site of all the transcripts in the human genome[262]. It computationally tests all known base pairing rules required to form RNA:DNA triplexes. The online version of the software was used by inputting the DNA sequence of the lncRNA of interest and querying against the whole genome (hg38). The output list of DNA binding motif hits were ranked according to the percentage of nuclei expressing the corresponding gene and the top 50 hits used as input to EnrichR to determine putative regulatory functions of the lncRNA under investigation.

2.4.12. Plots

All UMAP, heatmaps, violin, feature, and dot plots were produced with Seurat[200] and ggplot2[266]. Upset plots were generated using UpSetR[267].

2.5. 10X Genomics Visium validation of NS-Forest minimal marker genes

To validate the expression of genes from the snRNA-seq analyses, their expression profiles were examined in 10x Genomics Visium spatial transcriptomic datasets generated for a subset of samples. The genes chosen included selected marker genes from the NS-Forest analysis to validate their cell type-specific expression patterns.

2.5.1. Generation of 10X Genomics Visium spatial transcriptomic datasets

10X Genomics Visium spatial transcriptomic datasets were generated together with Ruvimbo Mishi (Masters student in the Hockman lab) using tissue originating from a subset of the same individuals used to generate the snRNA-seq datasets namely: the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old. The pipeline was carried out using the 10X Genomics Visium Spatial Gene Expression Kits according to manufacturer's protocols (User Guide Rev D). Briefly, 10 µm cryosections from OCT-embedded samples (2 sections per

sample) were placed on the capture areas of the spatial gene expression slide. The tissue was permeabilised for 12 minutes after reverse transcription was performed to generate barcoded cDNA libraries. The cDNA was amplified using 20 PCR cycles and libraries were constructed which included a size selection step. The quality of the cDNA was assessed before and after library preparation (Supp Fig 2.4-2.5, Supp Table 2.1). The libraries were sequenced by Novogene (Singapore) on the NovaSeq system using a NovaSeq High Output v.2.5 kit (150 cycles).

2.5.2. Pre-processing Visium data

Pre-processing of the Visium data was performed by Ruvimbo Mishi (Code availability, script 28). Briefly, Fastq file reads were aligned to a human reference transcriptome (GRCh38) using the count function from the 10X Genomics Space Ranger v1.3.0 software. The filtered gene barcode matrix outputs of Space Ranger for each sample were imported into R as a Seurat[202] object and a Visium anndata python object was created from the Seurat object (Code availability, script 29). Cell-level filtering was performed according to the following parameters nUMI > 500, nGene > 250, log10GenesPerUMI > 0.8 and mitoRatio < 0.2. Additionally, gene-level filtering was performed to remove genes that had zero counts in all cells, remove genes expressed in fewer than 3 cells, and remove mitochondrial genes from the gene by cell counts matrix.

2.5.3. Cell2Location to visualise gene expression in specific cell types

The Cell2Location[178] package was optimised to generate plots showing the estimated levels of expression and spatial location of genes of interest within specific cell types (Code availability, script 30-31). Briefly, the 23 integrated snRNA-seq datasets which were annotated using the Allen Institute's MTG snRNA-seq dataset served as a reference dataset. This was filtered to include counts > 0 in at least 3% of cells, mean expression > 1.12, and genes expressed in at least 5 cells. Subsequently, a negative binomial regression model was trained to estimate the expression of every gene in every cell type of this reference dataset using 250 epochs and adjusting for sample-specific variation. Spatial mapping was then performed to estimate the abundance of each cell type in the query Visium datasets. Vistoseg[268] was first used to determine the number of cells per location for input into cell2location. The datasets were then filtered to remove mitochondrial genes after which a model was prepared specifying two parameters, N_cells_per_location = 12 and detection_alpha =20. The model was subsequently trained using 2000 epochs. Lastly, the expression level of every gene at every spatial location in each cell type in the spatial data was estimated by applying a method adapted from Cable et al. (2022)[269] which uses the posterior distribution of cell-type specific expression instead of only using point estimates. The expression of the selected genes was then plotted for a subset of samples (DH2=4-year old, DH1a = 15-year-old B1, DH3= 15-year-old B2, DH4a=31-year-old) using Cell2Location's plot_genes_per_cell_type function. The expression profiles were visualised in a subset of relevant cell types namely Oligo L1-6 OPALIN, OPC L1-6 PDGFRA, Astro L1-6 FGFR3 SLC14A1,

Micro L1-3 TYROBP, Exc L2 LAMP5 LTK, Exc L4-6 RORB SEMA3E, Exc L5-6 FEZF2 ABO, Exc L4-6 FEZF2 IL26, Inh L2-4 PVALB WFDC2, and Inh L3-5 SST ADGRG6.

Chapter 3: Results

3.1. Pre-processing and quality control procedure for the raw 10X Genomics snRNA-seq datasets

A total of 23 snRNA-seq datasets from 12 different individuals were used in this study representing both biological and technical replicates (Table 2.1). Pre-filtering of the raw datasets was performed using Cell Ranger to remove barcodes that likely represent empty GEMs containing ambient RNA from lysed or dead cells. The sequencing saturation, which measures the proportion of total library complexity sequenced[134], varied from 27.4% in the 9-year-old T1 dataset to 79.1% in the 24-year-old T1 dataset (Supp Table 2.2). Notably, the publicly available datasets had a greater average sequencing saturation (68.0%) compared to the datasets generated in our laboratory (37.3%) (Supp Table 2.2). Additionally, the publicly available datasets had a greater average number of reads per nucleus (31738) compared to our datasets (17565) (Supp Table 2.2). However, the average number of genes detected per nucleus was slightly lower in the publicly available datasets (1285) compared to the new datasets (1470) (Supp Table 2.2).

Following initial filtering by Cell Ranger, additional quality control measures were performed using Seurat together with three doublet removal tools. Doublets are technical artefacts in single cell datasets that arise due to two nuclei getting captured together with a single GEM and thus the transcripts erroneously appear to originate from one nucleus. While other multiplets are possible they are rare compared to doublets so I focused on identifying and removing these[195]. To select barcodes for removal that are likely doublets, I chose to take the intersection of those nuclei called as doublets by DoubletDecon and Scrublet, as well as all nuclei called as doublets by DoubletFinder (Fig 3.1A, Supp Fig 3.1). This combination was chosen as a balance between the sensitivity and specificity of each tool since DoubletDecon has the highest sensitivity, Scrublet has the highest specificity and DoubletFinder has the highest accuracy overall[197]. Across the 23 datasets, an average of 1072 doublets were removed ranging from as few as 399 in the 24-year-old T1 dataset to 1876 in the 15-year-old T1 B2 dataset (Supp Fig 3.2A, Supp Table 3.1). The average doublet rate (percentage of total barcodes confidently called as doublets) was 12.62% with a minimum of 10.93% in the 50-year-old T1 dataset and a maximum of 15.97% in the 5-year-old T1 dataset (Supp Fig 3.2B, Supp Table 3.1-3.2).

After doublet removal, Seurat was used to compute additional quality control metrics for the remaining nuclei which were further filtered to retain those with a sufficiently high number of transcripts (nUMI > 500), a sufficiently high number of expressed genes (nGene > 250), a high complexity as indicated by the ratio of the number of genes expressed to the number of UMIs (log10 nGene/nUMI > 0.8), and a sufficiently low ratio of mitochondrial to total reads (mitoRatio <0.2) indicative of uncontaminated nuclei (Supp Fig 3.2C-F). Following filtering to remove poor quality barcodes, a total of 176 012 nuclei remained ranging from 2878 in the 24-year-old T1 dataset to 12516 in the 15-year-old T2 B1 dataset

(Fig 3.1B, Supp Table 3.2). The average number of barcodes per sample was 8390 before filtering and 7653 after filtering with a total of 16 951 nuclei removed (Supp Table 3.2). For the publicly available datasets, the average number of barcodes per sample was 5174 before filtering and 4535 after filtering (Supp Table 3.2). For the new datasets the average number of barcodes per sample was 9797 before filtering and 9017 after filtering (Supp Table 3.2). Overall, the filtered data had an average of 4101 UMIs and 1837 expressed genes per nucleus (Table 3.1). Furthermore, the average log10 ratio of number of expressed genes to number of transcripts was 0.93 and the average mitoRatio (number of non-mitochondrial to mitochondrial genes expressed) was 0.004 (Table 3.1, Supp Fig 3.2E-F). While the distribution of the number of transcripts (UMIs) present and genes expressed per nucleus was relatively similar across the datasets, the median number of UMIs and genes per nucleus for the 4-year-old and 26-year-old was higher than other datasets as these two samples were in the same batch and sequenced to a greater depth (Fig 3.1C-D, Table 2.1).

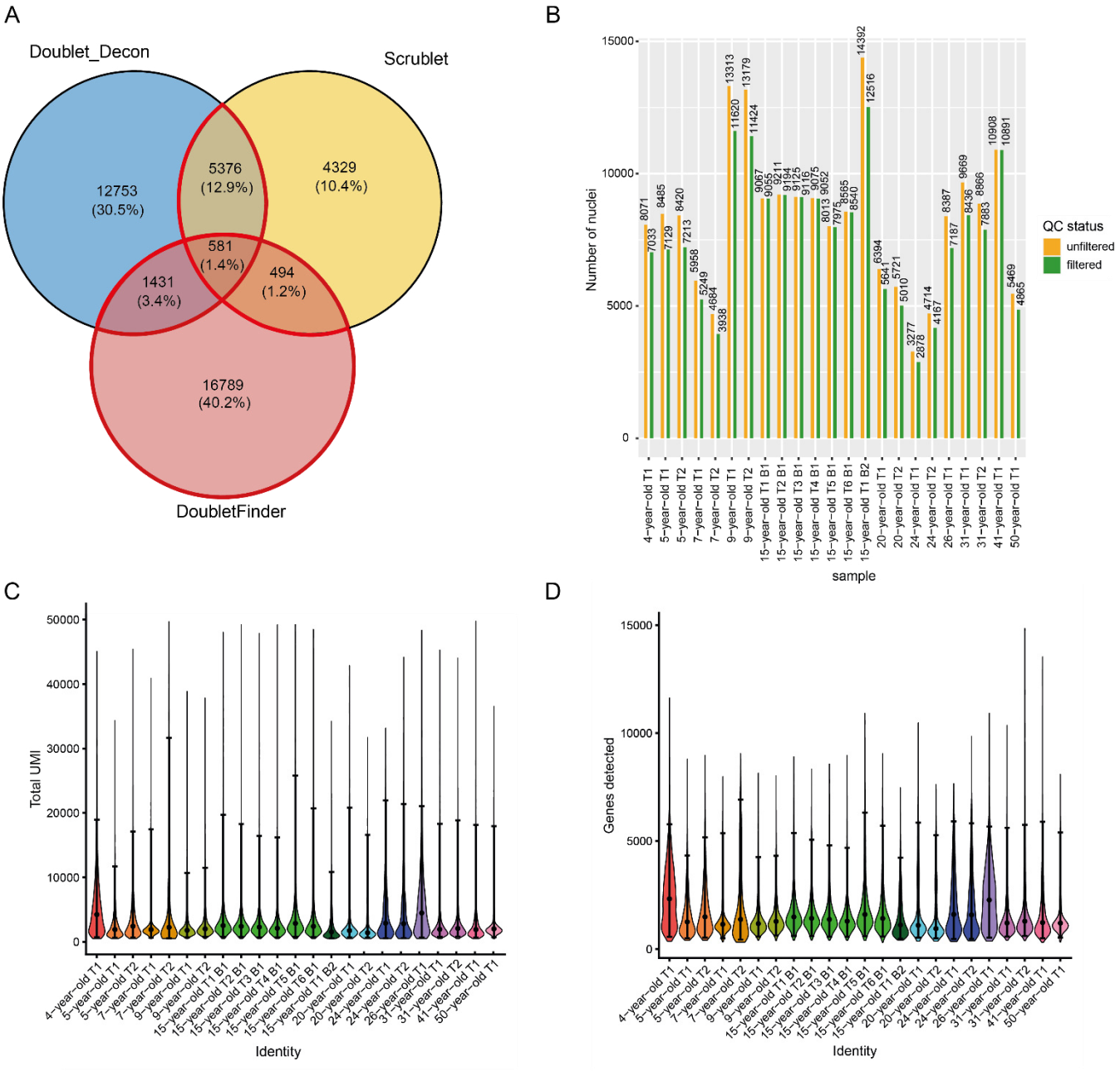


Figure 3.1. Cell and gene-level filtering measures to obtain high quality data. (A) Number of doublets identified across all 23 datasets by DoubletDecon, DoubletFinder, and Scrublet. Red outline indicates the subset of barcodes called as doublets that were removed from the cell x gene matrix. (B) Total number of nuclei per sample before (yellow) and after filtering (green). (C) Violin plots showing the number of unique molecular identifiers (UMIs) and (D) number of genes detected per nucleus per sample after filtering. Error bars represent mean +- SEM.

Table 3.1. Summary of average QC metrics across nuclei for each sample post filtering. Several measures for quality control were evaluated on a per sample basis including the average number of transcripts per nucleus (nUMI), the average number of genes detected per nucleus (nGene), the logarithm of the ratio of number of genes to transcripts (log10 Genes per UMI) observed, and the ratio of mitochondrial genes to total genes detected (mitoRatio).

sample	nUMI	nGene	log10 Genes per UMI	mitoRatio
4-year-old T1	5646	2522	0.93	0.0039
5-year-old T1	2979	1587	0.94	0.0014
5-year-old T2	4019	1873	0.93	0.0015
7-year-old T1	3463	1587	0.93	0.0033
7-year-old T2	6084	2134	0.93	0.0170
9-year-old T1	2627	1473	0.94	0.0016
9-year-old T2	2876	1556	0.94	0.0015
15-year-old T1 B1	4043	1808	0.93	0.0024
15-year-old T2 B1	3787	1732	0.93	0.0028
15-year-old T3 B1	3499	1654	0.93	0.0028
15-year-old T4 B1	3291	1573	0.93	0.0033
15-year-old T5 B1	5208	2105	0.92	0.0015
15-year-old T6 B1	4249	1875	0.93	0.0015
15-year-old T1 B2	2654	1505	0.95	0.0105
20-year-old T1	4572	1919	0.93	0.0024
20-year-old T2	3664	1700	0.93	0.0027
24-year-old T1	5264	2108	0.92	0.0061
24-year-old T2	5103	2052	0.92	0.0059
26-year-old T1	6102	2470	0.92	0.0022
31-year-old T1	3730	1718	0.93	0.0013
31-year-old T2	3894	1811	0.93	0.0015
41-year-old T1	3915	1818	0.94	0.0136
50-year-old T1	3648	1663	0.93	0.0031
Average	**4101**	**1837**	**0.93**	**0.0041**

3.2. Data integration, clustering, and annotation

Following quality control, the gene counts of each dataset were normalized and scaled using the Single Cell Transform[202] method to remove variability due to technical factors such as sequencing depth while preserving biological heterogeneity. An evaluation of the factors driving variation between nuclei revealed very few differences due to cell cycle

phase (Supp Fig 3.3A). On the other hand, mitoRatio appeared to be a notable source of variation between nuclei, with most nuclei having a low mitochondrial fraction (Supp Fig 3.3B).

Following normalisation and scaling, data integration was performed to align similar cell types across the 23 datasets (Fig 3.2A). This made an improvement on the nuclei alignment seen prior to integration (Supp Fig 3.3C). However, for largest group of nuclei (subsequently annotated as oligodendrocytes), there were several datasets whose nuclei were positioned separately from the rest and were not completely integrated suggesting a possible batch effect (Fig 3.2A). These outlying nuclei originated from technical replicates of just two of the samples (9-year-old T1, 9-year-old T2, 31-year-old T1, and 31-year-old T2) (Table 2.1) and were only observed after integration, being limited to the oligodendrocyte cluster (Fig 3.2A-B, Supp Fig 3.3C).

Following sample integration, dimensionality reduction and clustering analysis was performed which yielded 40 clusters (Supp Fig 3.3D). A combination of automated and manual annotation methods were used to classify the nuclei into various cell types. The automated annotation tools, scCATCH[217] and SCSA[218], revealed putative annotations for certain clusters, ambiguous annotations for other clusters, or no annotation (Supp Table 3.3). Likewise, violin plots of previously described cell type markers showed specificity of some markers for distinct clusters, however other markers appeared to be expressed in diverse clusters instead of localizing to a single or a few clusters (Supp Fig 3.4). For example, the expression of the astrocytic marker, *GFAP*, was limited to five clusters (9, 31, 36, 37 and 39) whereas the oligodendrocyte marker, *PLP1*, was expressed in all clusters, albeit at low levels in most clusters (Supp Fig 3.4C-D).

On the other hand, Seurat's label transfer method[205] unambiguously classified each nucleus into 1 of 54 different cell types using the Allen Brain Atlas human middle temporal gyrus (MTG) as the reference dataset[109] (Fig 3.2B). Of the 75 reference cell types, there were 21 which were absent from our dataset, including Exc L2-4 LINC00507 GLP2R, Exc L5-6 SLC17A7 IL15, Inh L1-2 LAMP5 DBP, and Inh L5-6 GAD1 GLP1R (Supp Table 3.4). To validate the high-resolution annotations, the expression of cell type-specific marker genes from Hodge et al. (2019)[109] was assessed across the clusters. This showed that each cell type-specific marker corresponded very clearly to the expected cell type (Fig 3.2C). To complement this approach, a correlation analysis was performed to compare the transcriptomic similarity of each of our annotated cell types to each of the reference MTG dataset cell types based on the median expression of cell type-specific markers (Fig 3.2D). The nonneuronal cell types in our dataset (including astrocytes, endothelial cells, microglia, OPCs, and oligodendrocytes) showed high correlation with the corresponding reference cell type as well as high specificity (low correlation with other cell types) (Fig 3.2D). On the other hand, the expression profiles of the inhibitory and excitatory neuronal subtypes correlated with multiple subtypes instead of showing a distinct correlation with a single subtype (Fig 3.2D). Nevertheless, the excitatory subtypes showed specificity for other excitatory subtypes while the inhibitory subtypes showed specificity for other inhibitory

subtypes. Additionally, most neuronal subtypes showed slightly higher correlation scores with their corresponding reference cell type compared to non-target subtypes as indicated by the diagonal line across the correlation plot (Fig 3.2D). Overall, based on the abovementioned annotation methods, a consensus of 7 broad cell types was determined (Supp Fig 3.5A, Supp Table 3.5). An assessment of the number of transcripts (UMIs) and the number of genes per nucleus per cell type revealed neuronal clusters to have a greater number of UMIs and genes expressed on average compared to non-neuronal cells (Fig 3.2E). Furthermore, excitatory neurons had a greater number of UMIs and genes expressed than inhibitory neurons (Fig 3.2E).

The technical replicates were collapsed for downstream analysis such that comparisons could be made between the 12 samples. The Oligo L1-6 OPALIN cluster had the greatest number of nuclei compared to the other clusters with 60 723 nuclei in total (Fig3.2C) and this observation was consistent across most samples except for the 4-year-old, 15-year-old B2, 24-year-old, and 26-year-old which had the highest number of nuclei in the Exc L5-6 THEMIS C1QL3, Exc L4-6 FEZF2 IL26, Micro L1-3 TYROBP, and Exc L2-3 LINC00507 FREM3 clusters, respectively (Fig 3.2F, Supp Table 3.6). After accounting for the total number of barcodes sequenced per sample, there appeared to be an increasing number of oligodendrocytes during the early postnatal years (ages 4 to 9) while the proportion of other clusters did not show such a clear trend with age (Fig 3.2F). A comparison of adults and paediatrics showed an increase in the proportion of Astro L1−2 FGFR3 GFAP, Astro L1−6 FGFR3 SLC14A1, and Exc L4−6 FEZF2 IL26 in paediatrics versus adults whereas the proportion of Oligo L1−6 OPALIN nuclei was slightly higher in adults versus paediatrics (Supp Fig 3.5B). However, overall, the cell composition of these two groups was reassuringly similar. Markedly, there appeared to be high variability in cell composition between technical replicates of the 5-year-old and 7-year-old whilst the other samples with technical replicates (9-year-old, 15-year-old B1, 20-year-old, 24-year-old, and 31-year-old) showed high degrees of similarity in cell composition between their replicates (Supp Fig 3.5C). Most notably, the proportion of Oligo L1−6 OPALIN nuclei in the 7-year-old T1 replicate was considerably greater than that of 7-year-old T2 whereas the proportion of Exc L2 LAMP5 LTK and Micro L1−3 TYROBP nuclei was considerably smaller (Supp Fig 3.5C).

The number of nuclei per MTG high resolution cluster per sample ranged from 0 to as many as 23 662 nuclei seen in the Oligo L1-6 OPALIN cluster of the 15-year-old B1 sample (Supp Table 3.6, Supp Fig 3.5D). Despite the 15-year-old B1 sample contributing more than double the number of nuclei compared to the other samples (n = 52932), this sample lacked nuclei for several cell sub-types, including several inhibitory neuron sub-types such as Inh L2-5 PVALB SCUBE3, Inh L1-3 VIP CCDC184, Inh L3-5 SST ADGRG6, and Inh L2-4 PVALB WFDC2 which the 5-year-old contributed the most nuclei to (Supp Table 3.6, Supp Fig 3.5D). Of note, there were several cell sub-types which lacked a contribution from some of the samples, including six inhibitory neuron sub-types (Inh L3-6 VIP HS3ST3A1, Inh L2-6 VIP QPCT, Inh L4-6 SST GXYLT2, Inh L1-2 VIP PCDH20, Inh L2-4 VIP SPAG17, Inh L2-5 VIP TYR) and three excitatory neuron sub-types (Exc L3-5 RORB FILIP1L, Exc L3-5 RORB TWIST2, Exc

L5-6 THEMIS FGF10) (Supp Table 3.6). Of these, five inhibitory neuron sub-types (Inh L2-6 VIP QPCT, Inh L4-6 SST GXYLT2, Inh L1-2 VIP PCDH20, Inh L2-5 VIP TYR, Inh L2-4 VIP SPAG17) and two excitatory neuron subtypes (Exc L3-5 RORB TWIST2 and Exc L5-6 THEMIS FGF10) all had fewer than 50 nuclei across all samples (Supp Table 3.6). The number of cell types annotated per sample ranged from 45 in the 50-year-old to 54 in the 15-year-old B1 (Supp Table 3.6).

In summary, through the label transfer approach, I was able to confirm the presence of many cell sub-types from the published MTG atlas. Subsequent validation further confirmed that they expressed the expected marker genes. However, not all 54 sub-types were found in all the samples.

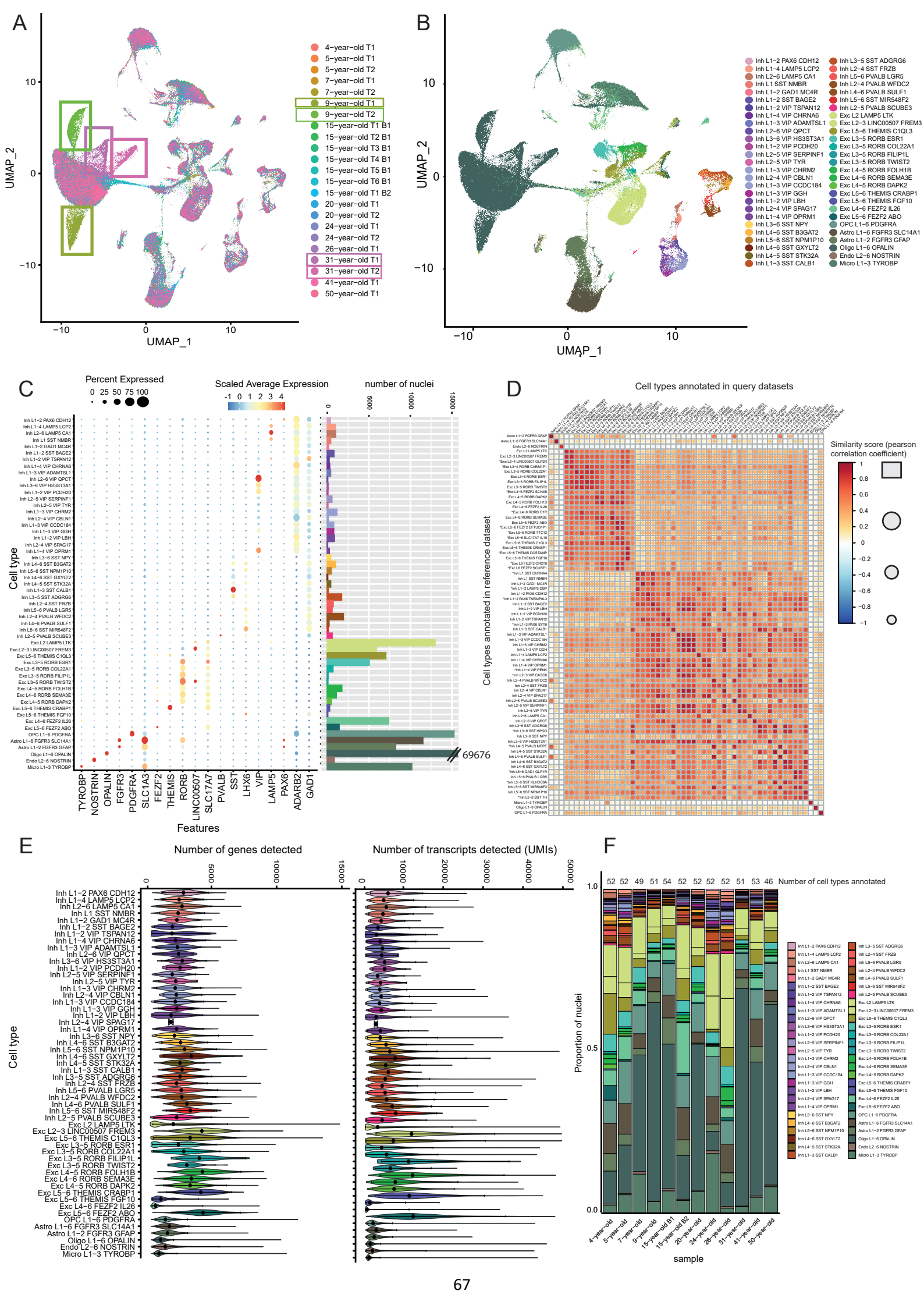

A
UMAP_2
UMAP_1
4−year−old T1
5−year−old T1
5−year−old T2
7−year−old T1
7−year−old T2
9−year−old T1
9−year−old T2
15−year−old T1 B1
15−year−old T2 B1
15−year−old T3 B1
15−year−old T4 B1
15−year−old T5 B1
15−year−old T6 B1
15−year−old T1 B2
20−year−old T1
20−year−old T2
24−year−old T1
24−year−old T2
26−year−old T1
31−year−old T1
31−year−old T2
41−year−old T1
50−year−old T1
B
UMAP_2
UMAP_1
Inh L1−2 PAX6 CDH12
Inh L1−4 LAMP5 LCP2
Inh L2−6 LAMP5 CA1
Inh L1 SST NMBR
Inh L1−2 GAD1 MC4R
Inh L1−2 SST BAGE2
Inh L1−2 VIP TSPAN12
Inh L1−4 VIP CHRNA6
Inh L1−3 VIP ADAMTSL1
Inh L2−6 VIP QPCT
Inh L3−6 VIP HS3ST3A1
Inh L1−2 VIP PCDH20
Inh L2−5 VIP SERPINF1
Inh L2−5 VIP TYR
Inh L1−3 VIP CHRM2
Inh L2−4 VIP CBLN1
Inh L1−3 VIP CCDC184
Inh L1−3 VIP GGH
Inh L1−2 VIP LBH
Inh L2−4 VIP SPAG17
Inh L1−4 VIP OPRM1
Inh L3−6 SST NPY
Inh L4−6 SST B3GAT2
Inh L5−6 SST NPM1P10
Inh L4−6 SST GXYLT2
Inh L4−5 SST STK32A
Inh L1−3 SST CALB1
Inh L3−5 SST ADGRG6
Inh L2−4 SST FRZB
Inh L5−6 PVALB LGR5
Inh L2−4 PVALB WFDC2
Inh L4−6 PVALB SULF1
Inh L5−6 SST MIR548F2
Inh L2−5 PVALB SCUBE3
Exc L2 LAMP5 LTK
Exc L2−3 LINC00507 FREM3
Exc L5−6 THEMIS C1QL3
Exc L3−5 RORB ESR1
Exc L3−5 RORB COL22A1
Exc L3−5 RORB FILIP1L
Exc L3−5 RORB TWIST2
Exc L4−5 RORB FOLH1B
Exc L4−6 RORB SEMA3E
Exc L4−5 RORB DAPK2
Exc L5−6 THEMIS CRABP1
Exc L5−6 THEMIS FGF10
Exc L4−6 FEZF2 IL26
Exc L5−6 FEZF2 ABO
OPC L1−6 PDGFRA
Astro L1−6 FGFR3 SLC14A1
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Micro L1−3 TYROBP
C
Percent Expressed
0 25 50 75 100
Scaled Average Expression
-1 0 1 2 3 4
number of nuclei
0
5000
10000
15000
69676
Cell type
Features
TYROBP
NOSTRIN
OPALIN
FGFR3
PDGFRA
SLC1A3
FEZF2
THEMIS
RORB
LINC00507
SLC17A7
PVALB
SST
LHX6
VIP
LAMP5
PAX6
ADARB2
GAD1
D
Cell types annotated in query datasets
Cell types annotated in reference dataset
Similarity score (pearson correlation coefficient)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
E
Number of genes detected
0
5000
10000
15000
Number of transcripts detected (UMIs)
0
10000
20000
30000
40000
50000
Cell type
F
52 52 49 51 54 52 52 52 52 51 53 46 Number of cell types annotated
1.0
0.5
0.0
Proportion of nuclei
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old
sample

Figure 3.2. Annotation of nuclei by label transfer identifies 54 cortical subtypes across the 23 datasets. (A) UMAP plot after data integration shows nuclei across the different datasets aligned, with the exception of four groups of nuclei which clustered separately (outlined by coloured rectangles). (B) Annotated UMAP plot for the 23 merged datasets identifies 34 inhibitory, 14 excitatory, and 6 non-neuronal populations using the Allen Brain Institute's MTG dataset served as the reference for the label transfer method. (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 across all 23 datasets (right). (D) Similarity plot showing the Pearson correlation scores assessing similarity between the annotated cell types in our dataset (x axis) and the MTG reference dataset (y axis) based on the median expression of cell type-specific marker genes. The cell type-specific marker genes were chosen based on having a beta score > 0.5. Similarity is indicated by a colour scale (blue-yellow-red) and by shape (small circle-medium circle-large circle-square). (E) 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. (F) Stacked barplot showing the proportion of nuclei per cell type (y axis) for each sample (x axis) out of the total number of nuclei for each sample. The colour scheme for the cell types is in accordance with the MTG dataset taxonomy[109].

3.3. NS-Forest minimal marker gene combination analysis

In order to establish a standardized and scalable approach for defining cell types identified in scRNA-seq studies it has previously been proposed to use the minimum combination of gene markers that can classify a cell type and distinguish it from other cell types[219]. Towards achieving this, Aevermann et al. (2021) developed a machine learning tool called Necessary and Sufficient Forest (NS-Forest) which attempts to identify marker genes that are simultaneously highly expressed in all individual cells of a population and are not expressed in other cell types – thus having strong classification power. Here I applied the NS-Forest algorithm to each of the 12 datasets. This was done to assess the variability in minimal marker genes identified for each of the 54 cell sub-types across the different samples and to identify any marker genes distinguishing the paediatric and adult datasets.

Across the 54 high resolution MTG clusters and 12 samples, a total of 761 unique marker genes were determined necessary to distinguish cell types with an average of 122.5 markers per sample (Fig 3.3A, Supp Table 3.7). The average number of marker genes necessary to classify a cell type was 2.39 with a minimum of 1 and a maximum of 5 markers (Supp Fig 3.6, Supp Table 3.8). The log-transformed average normalized expression of the markers per nucleus within the cell type of interest across all samples was 2.902 with a minimum of 0.0076 for *LINC02017* in Inh L1-2 VIP TSPAN12 and a maximum of 84.84 for *CCK* also in Inh L1-2 VIP TSPAN12 (Supp Table 3.9). The average F-beta score (measure of classification power) for the marker gene combinations per cell type was 0.67 (Fig 3.3B) whilst the average binary expression score per cell type was 0.96 (Fig 3.3C). The F-beta score assesses combinations of marker genes with a score of 1 indicating optimal discriminative power of a given combination of marker genes. On the other hand, the binary expression score assesses individual marker genes with a score of 1 indicating high levels of expression in most of the nuclei in the target cluster and no expression in off-target clusters. Notably, paediatric samples had a slightly lower median average F-beta score for the marker gene combinations per cell type compared to adults (0.63 vs 0.69) (Fig 3.3B) as well as a negligibly

lower median average binary expression score for the minimal marker genes (0.96 vs 0.963) (Fig 3.3C).

The number of marker genes that corresponded to markers found in Hodge et al.(2019) or Aevermann et al.(2021) were 7 and 28, respectively, with 3 markers found in both Hodge and Aevermann (Fig 3.3A, Supp Fig 3.6B-C, Supp Table 3.10). For example, *MYO5B, KIT*, and *DACH2* were common marker genes in our data and Aevermann et al. (2021) for the Inh L2-4 SST FRZ, Inh L1-4 LAMP5 LCP2, and Inh L1-3 VIP ADAMTSL1 clusters, respectively (Supp Fig 3.6C). *STK32A* agreed as a marker gene in our data, Aevermann et al. (2021), and Hodge et al. (2019) for Inh L4-5 SST STK32A (Supp Fig 3.6B-C). Additionally, there were several well-known markers genes among the list of minimal markers, including *APBB1IP* for Micro L1-3 TYROBP[270], *PDGFRA* for OPC L1-6 PDGFRA[271], *ST18* for Oligo L1-6 OPALIN[108], and *GFAP* for Astro L1-2 FGFR3 GFAP[272] – many of which were shared with either Aevermann or Hodge (Fig 3.3A, Supp Fig 3.6B-C). A total of 190 genes were identified as a marker for at least two cell types (across the 12 samples) (Supp Table 3.11). To validate the ability of the NS-Forest marker genes to distinguish different cell types, the data integration method was repeated using either the marker genes (Fig 3.3D) or the equivalent number of a random set of genes as anchors as opposed to the standard shared highly variable genes (Fig 3.3E). The resulting UMAP plot using the NS-Forest genes more closely resembled the original UMAP plot compared to when a random set of genes was used (compare Fig3.2A, Fig 3.3D, and Fig 3.3E).

With regards to the gene class of the markers identified, 443 marker genes were coding (58.2%) while 318 were non-coding (41.8%)(Fig 3.3A). Exc L3-5 RORB FILIP1L in particular had a large proportion of non-coding minimal markers with 21 non-coding and 1 coding gene identified as marker genes (Fig 3.3F, Supp Table 3.12). Thus, the ratio of non-coding to coding genes identified was 21, which exceeded the ratio computed for all other cell types by at least 7 times (Supp Table 3.12). Similarly, Exc L4-5 RORB FOLH1B and Exc L3-5 RORB TWIST2 also had a larger number of non-coding than coding marker genes (Fig 3.3F, Supp Table 3.12). Markedly, for both Exc L3-5 RORB FILIP1L and Exc L3-5 RORB TWIST2 all the markers identified (both coding and non-coding) were unique to a single sample (Supp Fig 3.7-3.9, Supp Table 3.13). However, there are numerous non-coding marker genes which were shared between multiple samples, including *AC046195.2* in Inh L5-6 PVALB LGR5, *LINC01344* in Inh L2-6 LAMP5 CA1, *AC137770.1* in Inh L1-4 LAMP5 LCP2, *AC021134.1* in Exc L4-5 RORB FOLH1B, and *CYP1B1-AS1* in Exc L2-3 LINC00507 FREM3 (Supp Fig 3.7-3.9, Supp Table 3.13). Overall, it appeared that the non-neuronal cell types had a lower non-coding to coding marker ratio than the neuronal cell types (Supp Table 3.12). For each of the samples, there was a greater number of minimal markers that were coding genes than non-coding genes, with a similar ratio observed across the samples (Fig 3.3G).

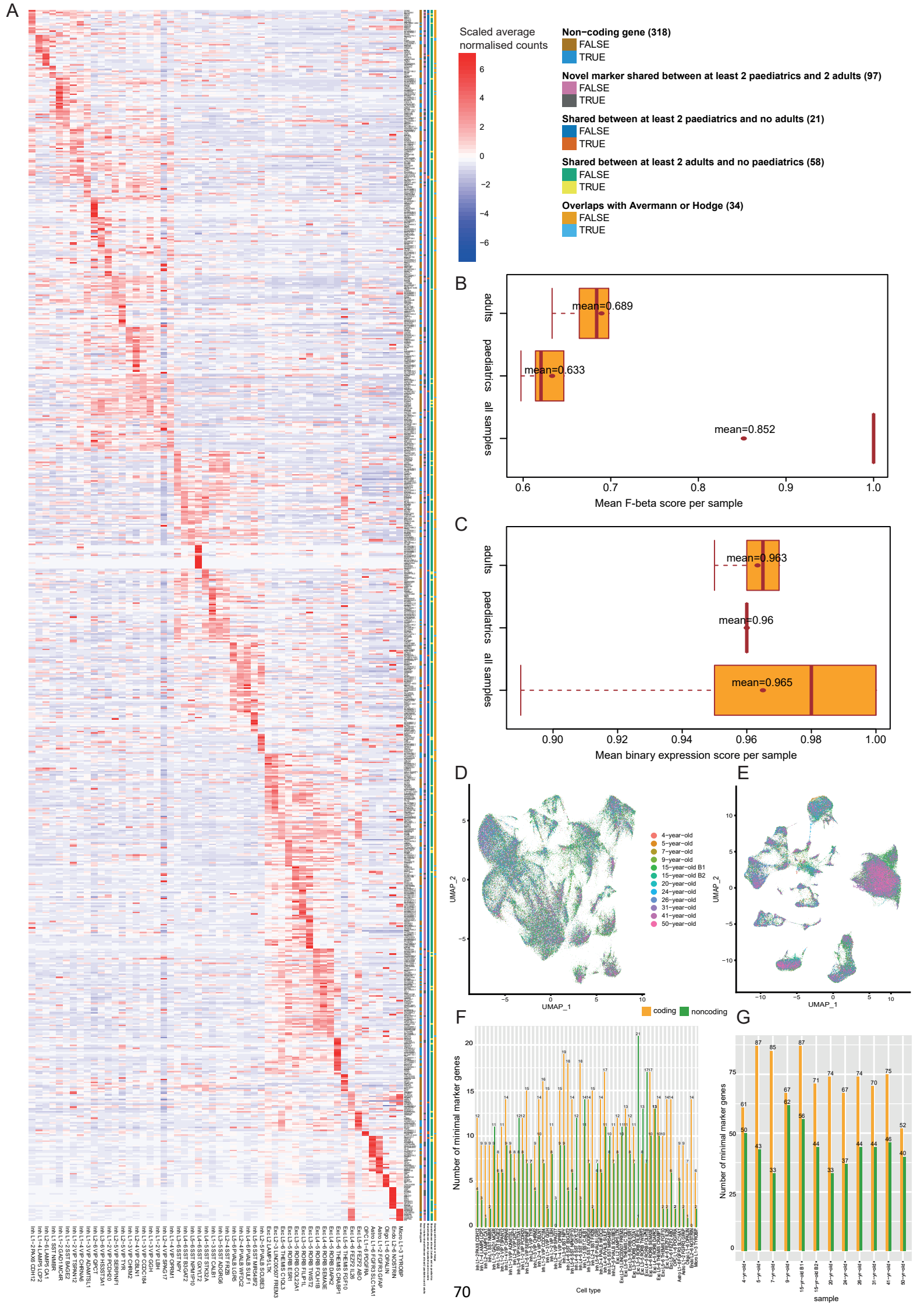

A
Scaled average normalised counts
6
4
2
0
−2
−4
−6
Non−coding gene (318)
FALSE
TRUE
Novel marker shared between at least 2 paediatrics and 2 adults (97)
FALSE
TRUE
Shared between at least 2 paediatrics and no adults (21)
FALSE
TRUE
Shared between at least 2 adults and no paediatrics (58)
FALSE
TRUE
Overlaps with Avermann or Hodge (34)
FALSE
TRUE
B
adults
paediatrics
all samples
mean=0.689
mean=0.633
mean=0.852
0.6
0.7
0.8
0.9
1.0
Mean F-beta score per sample
C
adults
paediatrics
all samples
mean=0.963
mean=0.96
mean=0.965
0.90
0.92
0.94
0.96
0.98
1.00
Mean binary expression score per sample
D
UMAP_2
UMAP_1
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old
E
UMAP_2
UMAP_1
coding
noncoding
F
Number of minimal marker genes
Cell type
G
Number of minimal marker genes
sample
4−yr−old
5−yr−old
7−yr−old
9−yr−old
15−yr−old B1
15−yr−old B2
20−yr−old
24−yr−old
26−yr−old
31−yr−old
41−yr−old
50−yr−old

Figure 3.3. NS-Forest identifies minimal marker genes distinguishing the cortical cell types in each of the 12 samples. (A) Heatmap showing the scaled average normalised expression counts of the 761 NS-Forest minimal marker genes (y-axis) identified across the 12 datasets for each of the 54 cortical cell types (x-axis). The marker genes are annotated according to their coding/noncoding status, the number of paediatric and/or adult samples expressing the gene, and the overlap of the markers with those from Hodge et al.(2019)[109] or Aevermann et al.(2021)[219] for the same cell type. (B) Average F-beta score for the marker gene combinations per cell type for all samples (bottom), paediatric samples (middle), or adult samples (top). The F-beta score assesses the classification power of combinations of marker genes based on an expression difference between the target cluster and off-target clusters. A score of 1 indicates high discriminative power of the markers while a score of 0 indicates low discriminative power. (C) Average binary expression score of the marker genes per cell type for all samples (bottom), paediatric samples (middle), or adult samples (top). The binary expression score assesses whether individual marker genes show binary expression with a score of 1 indicating high levels of expression in a large proportion of nuclei in the target cluster and no expression in off-target clusters while a score of 0 indicates high off-target expression. (D, E) Validation of the relevance of the NS-Forest markers in distinguishing different cell types by repeating the data integration method using either the marker genes (D) or the equivalent number of a random set of genes (E) as anchors. (F, G) Number of coding versus noncoding marker genes per cell type (F) and per sample (G). The coding status was determined using the list of unique marker genes per cell type and sample.

To assess the variability between the samples in terms of the markers identified for each cluster, upset plots were generated (Fig 3.4, Supp Fig 3.7-3.9). These revealed that most minimal markers were either unique to a sample or shared between two samples with fewer cases where a marker was shared between 3 or more samples (Fig 3.4, Supp Fig 3.7-3.9). As an exception to this observation, *SEMA3E* was shared between 10 of the samples in the Exc L5-6 FEZF2 ABO cluster (Fig 3.4A, Supp Table 3.13) whilst *AC004852.2* was shared between 10 samples in OPC L1-6 PDGFRA (Fig 3.4B, Supp Table 3.13), suggesting that these may represent novel consensus markers for these cell types. Other markers shared between multiple samples include *DDR2* (7 samples, Inh L1-2 PAX6 CDH12) and *EDNRA* (8 samples, Inh L4-5 SST STK32A) (Supp Table 3.13). Interestingly, there were several markers which were largely shared between paediatric samples and not adults and vice versa including *PTH2R* which was shared between the 5-year-old, 7-year-old, 9-year-old, and 15-year-old B1 in the Inh L1-3 VIP GGH cell type (Fig 3.4C, Supp Table 3.13) whereas *RSPO2* was shared between the 31-year-old, 41-year-old, and 50-year-old in the Inh L3-5 SST ADGRG6 cluster (Fig 3.4E, Supp Table 3.13). Additionally, a noncoding gene, *AC021134.1*, was shared between the 5-year-old, 7-year-old, 9-year-old, 15-year-old B2, and 20-year-old in the Exc L4-5 RORB FOLH1B cluster (Fig 3.4D, Supp Table 3.13) whilst *LINC00499* was shared between the 15-year-old-B1, 31-year-old, 41-year-old, and 50-year-old in Astro L1-6 FGFR3 SLC14A1 (Fig 3.4F, Supp Table 3.13). Unexpectedly, *LRMDA*, which is a known microglial marker gene, was identified as a marker for the Exc L4-6 FEZF2 IL26 cell type in the 9-year-old, 15-year-old B1, and 15-year-old B2 suggesting that the annotation for this cell type could be incorrect (Fig 3.4H, Supp Table 3.13). However, an examination of the similarity between our annotated cell types and the reference MTG cell types suggests that the Micro L1-3 TYROBP and Exc L4-6 FEZF2 IL26 cell types are transcriptionally distinct, since Exc L4-6 FEZF2 IL26 in our dataset is more similar to the equivalent Exc L4-6 FEZF2 IL26 population than it is to the Micro L1-3 TYROBP population in the reference MTG dataset (Fig 3.2D).

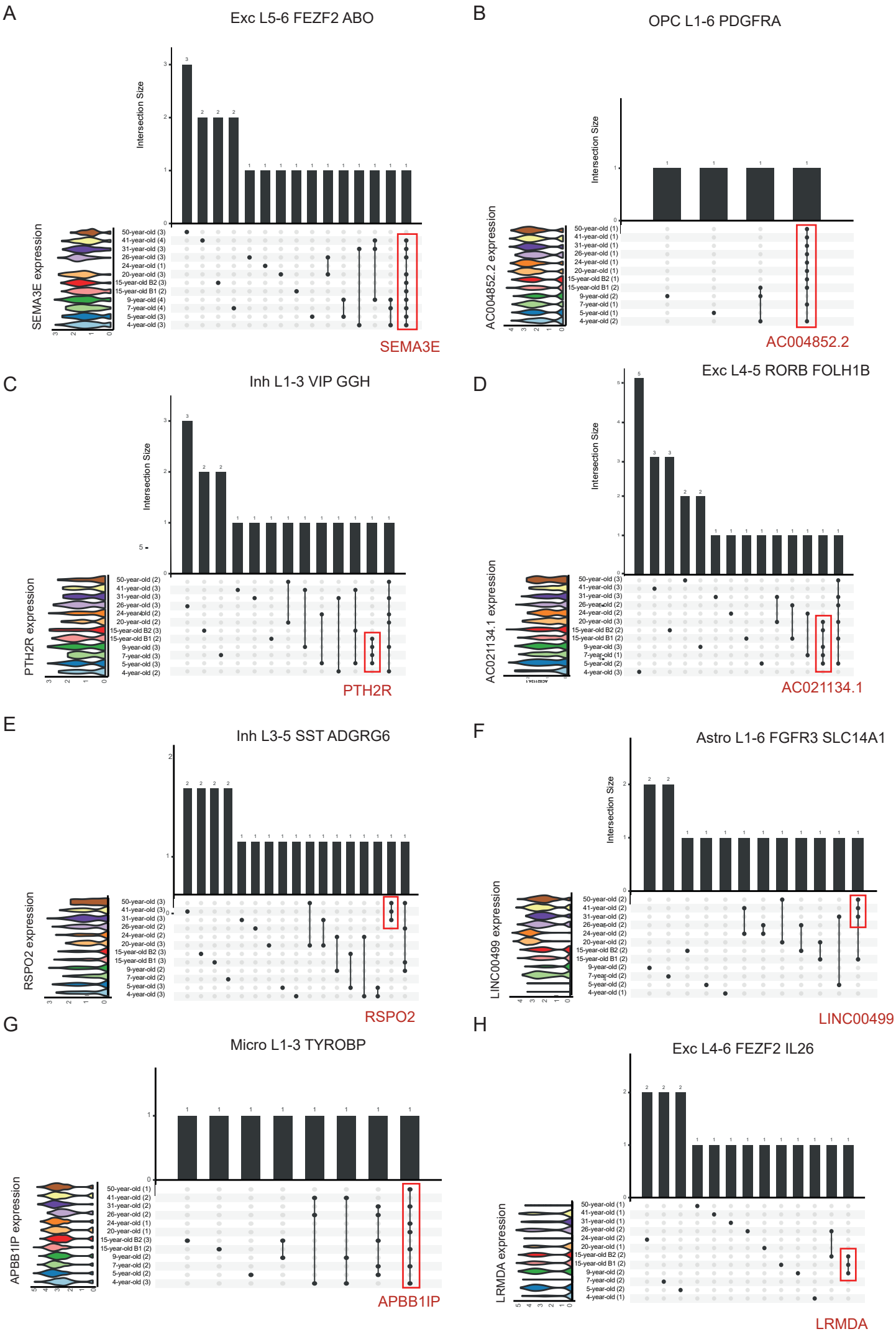

A
Exc L5-6 FEZF2 ABO
Intersection Size
SEMA3E expression
50-year-old (3)
41-year-old (4)
31-year-old (3)
26-year-old (3)
24-year-old (1)
20-year-old (3)
15-year-old B2 (3)
15-year-old B1 (2)
9-year-old (4)
7-year-old (4)
5-year-old (3)
4-year-old (3)
SEMA3E
B
OPC L1-6 PDGFRA
Intersection Size
AC004852.2 expression
50-year-old (1)
41-year-old (1)
31-year-old (1)
26-year-old (1)
24-year-old (1)
20-year-old (1)
15-year-old B2 (1)
15-year-old B1 (2)
9-year-old (2)
7-year-old (1)
5-year-old (1)
4-year-old (2)
AC004852.2
C
Inh L1-3 VIP GGH
Intersection Size
PTH2R expression
50-year-old (2)
41-year-old (3)
31-year-old (3)
26-year-old (3)
24-year-old (2)
20-year-old (2)
15-year-old B2 (3)
15-year-old B1 (2)
9-year-old (3)
7-year-old (3)
5-year-old (3)
4-year-old (2)
PTH2R
D
Exc L4-5 RORB FOLH1B
Intersection Size
AC021134.1 expression
50-year-old (3)
41-year-old (3)
31-year-old (3)
26-year-old (2)
24-year-old (2)
20-year-old (3)
15-year-old B2 (2)
15-year-old B1 (2)
9-year-old (3)
7-year-old (1)
5-year-old (2)
4-year-old (3)
AC021134.1
E
Inh L3-5 SST ADGRG6
RSPO2 expression
50-year-old (3)
41-year-old (3)
31-year-old (3)
26-year-old (2)
24-year-old (2)
20-year-old (3)
15-year-old B2 (3)
15-year-old B1 (3)
9-year-old (2)
7-year-old (2)
5-year-old (3)
4-year-old (3)
RSPO2
F
Astro L1-6 FGFR3 SLC14A1
Intersection Size
LINC00499 expression
50-year-old (2)
41-year-old (2)
31-year-old (2)
26-year-old (2)
24-year-old (2)
20-year-old (2)
15-year-old B2 (2)
15-year-old B1 (2)
9-year-old (2)
7-year-old (2)
5-year-old (2)
4-year-old (1)
LINC00499
G
Micro L1-3 TYROBP
APBB1IP expression
50-year-old (1)
41-year-old (2)
31-year-old (2)
26-year-old (2)
24-year-old (1)
20-year-old (1)
15-year-old B2 (3)
15-year-old B1 (2)
9-year-old (2)
7-year-old (2)
5-year-old (2)
4-year-old (3)
APBB1IP
H
Exc L4-6 FEZF2 IL26
LRMDA expression
50-year-old (1)
41-year-old (1)
31-year-old (1)
26-year-old (2)
24-year-old (2)
20-year-old (1)
15-year-old B2 (2)
15-year-old B1 (2)
9-year-old (2)
7-year-old (2)
5-year-old (2)
4-year-old (1)
LRMDA

Figure 3.4. Several NS-Forest minimal marker genes are shared between paediatric samples and not adults or vice versa. (A-H) Upset plots show the overlap of NS-Forest minimal marker genes between samples for a subset of cell types: Exc L5-6 FEZF2 ABO (A), OPC L1-6 PDGFRA (B), Inh L1-3 VIP GGH (C), Exc L4-5 RORB FOLH1B (D), Inh L3-5 SST ADGRG6 (E), Astro L1-6 FGFR3 SLC14A1 (F), Micro L1-3 TYROBP (G), and Exc L4-6 FEZF2 IL26 (H). Marker genes of interest are highlighted which are either shared between multiple samples, shared between paediatric and not adult samples, or shared between adult and not paediatric samples (red boxes). On the left of each plot the total number of markers per sample is indicated in brackets. Violin plots show the expression levels of the marker genes of interest across the samples.

3.4. Differential gene expression analysis between different age groups within specific cell types using DESeq2

To identify genes whose expression changes with age, the samples were grouped into five epochs (Materials and Methodology section 2.4.6 DE section). Principal component analysis on each cluster was used to reveal variables that contributed to inter-sample variation. This analysis revealed that for most clusters, samples primarily separated out according to the single cell chemistry platform used to generate the datasets (version 2 chemistry vs version 3 chemistry) (Supp Fig 3.10). Given this result, single-cell chemistry was included as a covariate in the regression formula to model its effect and adjust the p values accordingly. To assess the fit of the model for the data, plots of the dispersion estimates were generated for each cell type. This revealed a decrease in the dispersion estimates with increasing mean for each cell type and the data points followed the line of best-fit suggesting the model is a good fit for the data (Supp Fig 3.11).

A likelihood ratio test was performed using DESeq2[209] in order to compare the expression of genes between epochs for each of the MTG clusters separately. This analysis yielded 673 significantDEGs across 35 cell types, with the Micro L1-3 TYROBP cluster having the greatest number of DEGs at 104 genes (Fig 3.5A, Supp Table 3.14). The DEGs represent genes which showed a significant change in expression between at least two epochs with the p-values and adjusted p-values computed for all pairwise comparisons together as opposed to individual pairwise comparisons. The log2 Fold Changes for individual pairwise comparisons (e.g epoch 1 vs epoch 5) can be found in Extended Data 2.

There appeared to be a moderate correlation (r = 0.601) between the number of DEGs identified per cell type and the total number of nuclei analysed for that cell type (Fig 3.5B). This correlation score increased to 0.856 after removing the data point corresponding to the Oligodendrocyte population which was an outlier as it had more than 4 times the number of nuclei than any other cell type (Supp Fig 3.12A). Notably, there were 18 cell types (mostly inhibitory neuron subtypes) in which no DEGs were identified and 39 cell types in which fewer than 10 DEGs were identified (Fig 3.5A). The Inh L2-4 VIP SPAG17 cluster was removed from the analysis due to having too few nuclei in total (n = 2) with several epochs having no nuclei for this cell type (Fig3.5A, Supp Table 3.6). Numerous DEGs were expressed in a low percentage (< 20%) of nuclei in the population of interest (Supp Table 3.14).

An analysis of the intersection of DEGs between the clusters indicated that age-related changes in gene expression are largely cluster-specific, with 425 genes being unique to a

single cell type (Supp Fig 3.12B, Extended Data 3). The populations with the largest number of unique genes included Micro L1-3 TYROBP (91), Oligo L1-6 OPALIN (60), and Exc L3-5 RORB ESR1 (41) (Supp Fig 3.12B, Extended Data 3). Curiously, there was one gene, *ARL17B*, which was shared between 27 cell types (Supp Fig 3.12B, Supp Table 3.14). Within the excitatory neuron group, Exc L2 Lamp5 LTK and Exc L2−3 LINC00507 FREM3 were most similar with 9 DEGs shared exclusively between these two clusters (Fig 3.5C) out of a total 16 shared DEGs (i.e. the remaining 7 DEGs were also shared with other cell types) (Extended Data 3). This was followed by 4 genes shared exclusively between Exc L3−5 RORB ESR1 and Exc L4−6 RORB SEMA3E, as well as 4 genes between Exc L4−6 RORB SEMA3E and Exc L5−6 THEMIS CRABP1 (Fig 3.5C). Within the inhibitory neuron group, *AL162493.1* and *CHRM3-AS2* were shared between two subtypes, while *ARL17B* was shared between twelve subtypes (Fig 3.5D, Supp Table 3.14). Of the non-neuronal clusters, the two astrocyte clusters, Astro L1−6 FGFR3 SLC14A1 and Astro L1−2 FGFR3 GFAP, shared the greatest number of genes with 11 DEGs shared exclusively (Fig 3.5E) out of a total of 13 shared DEGs (Extended Data 3). This was followed by 3 genes shared exclusively between Micro L1−3 TYROBP and Oligo L1−6 OPALIN (6 shared in total); Micro L1−3 TYROBP and Astro L1−2 FGFR3 GFAP (5 shared in total); as well as Oligo L1−6 OPALIN and OPC L1−6 PDGFRA (8 shared in total) (Fig 3.5E, Extended Data 3). The number and proportion of coding versus non-coding DEGs per cell type was variable with no clear trend distinguishing the different cell types, however the majority of DEGs were coding genes (Fig 3.5F, Supp Table 3.14).

For each cell type, the DEGs were grouped into similar patterns of expression with age (Supp Fig 3.13, Supp Table 3.15). Considering the relatively low number of DEGs per pattern, patterns which showed a general increasing or decreasing trajectory in their level of expression were grouped together to increase the statistical power forGSEA. GSEA was performed using the GO Biological processes database, GTEx Aging signature database, and the DisGeNET disease database. The results of this analysis are summarised in Supplementary text 1 and Supplementary figures 3.14-3.17, with the top 10 terms from each database shown. Markedly, there were many terms which were not significantly enriched after adjusting for multiple testing. The full enrichment results are in Supp Table 3.16.

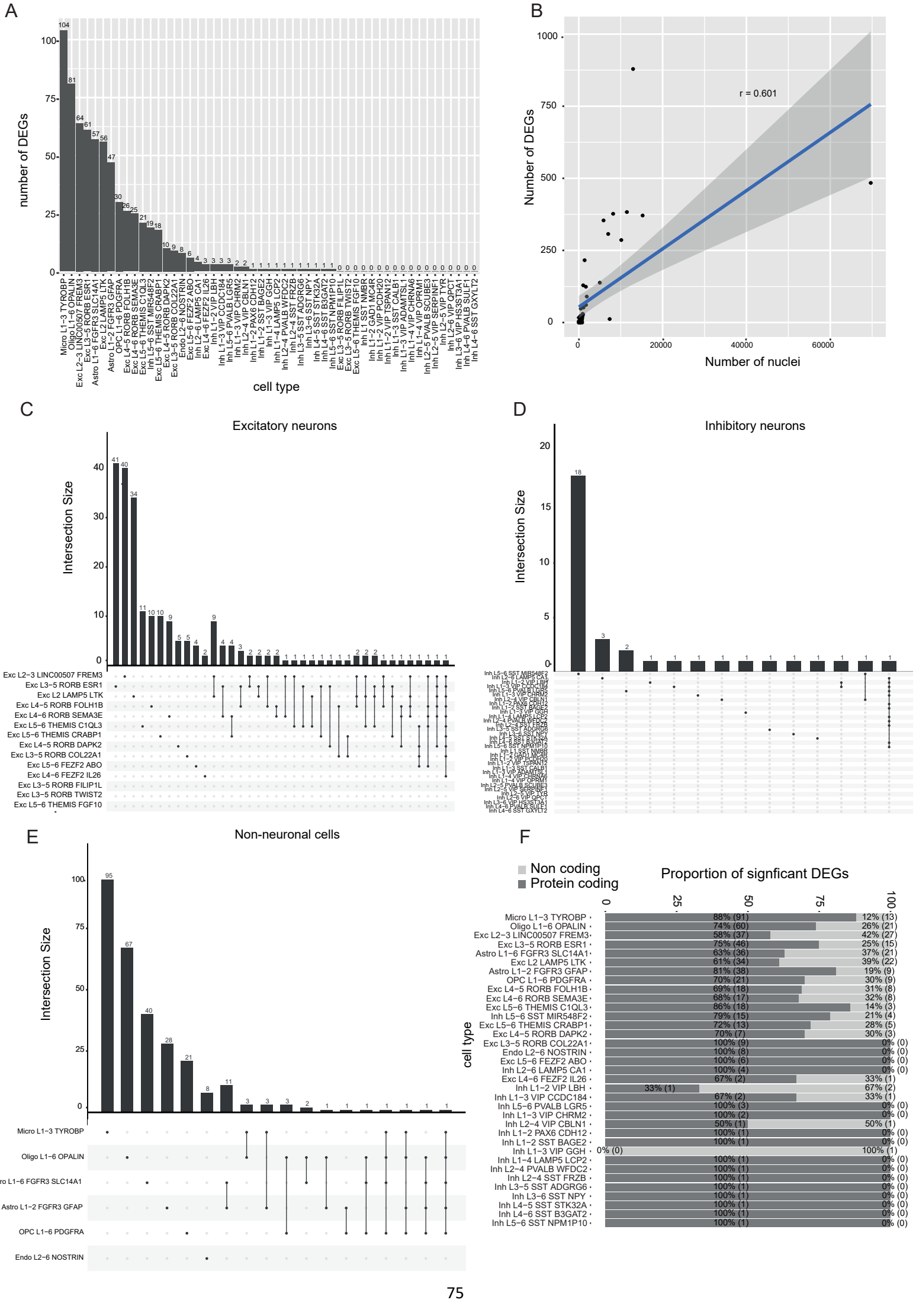

A
number of DEGs
cell type
B
Number of DEGs
Number of nuclei
r = 0.601
C
Excitatory neurons
Intersection Size
D
Inhibitory neurons
Intersection Size
E
Non-neuronal cells
Intersection Size
F
Non coding
Protein coding
Proportion of signficant DEGs
cell type

Figure 3.5. DESeq2 identifies age-dependent DEGs in neuronal and non-neuronal cell types. (A) Number of DEGs identified per cell type. Samples were grouped into five different age groups (epochs) and DEGs were determined as genes that showed a change in their level of expression with age between at least two epochs. (B) Correlation between the number of nuclei (x axis) and the number of DEGs (y axis) identified for each cell type based on Pearson's correlation. (C-E) Upset plots showing the overlap of DEGs between the excitatory neuron subtypes (C), the inhibitory neuron subtypes (D), and the nonneuronal subtypes (E). The size of the intersections is indicated above the bars. The overlapping genes represent only the unique intersections between any set of cell types. (F) Proportion of DEGs per cell type that are coding or non-coding genes. The absolute number of coding and non-coding DEGs per cell type are indicated in brackets.

3.5. Time-series Psupertime analysis to identify genes varying coherently with age within specific cell types

As a complementary approach to DESeq2, Psupertime[226] was used to identify temporally regulated genes across the time-series. This was performed for each cell type separately which yielded 15392 temporally regulated genes across 33 cell types (Fig 3.6A, Supp Fig 3.19, Supp Table 3.17). 21 cell types were excluded from this analysis due to having too few nuclei in at least one sample.

The largest number of temporally regulated genes was identified in Exc L2 LAMP5 LTK (2370 genes), followed by Astro L1−6 FGFR3 SLC14A1 (1922 genes), and Exc L2−3 LINC00507 FREM3 (1838 genes) (Fig 3.6A, Supp Table 3.17). As with the DESeq2 analysis, there was a moderate correlation between the number of temporally regulated genes identified and the number of nuclei per cell type (r = 0.55) (Fig 3.6B) which increased after removing the oligodendrocyte datapoint (r = 0.872) (Supp Fig 3.12C). The proportion of temporally regulated genes that were protein coding genes was at least 75% for each of the cell types (Fig 3.6D).

Similar to what was seen for DESeq2, an analysis of the overlap of temporally regulated genes between cell types showed the greatest intersection to be between Exc L2−3 LINC00507 FREM3 and Exc L2 LAMP5 LTK (120 genes shared exclusively out of 496 shared genes) (Fig 3.6C, Extended Data 4). This was followed by Exc L2 LAMP5 LTK and Astro L1−6 FGFR3 SLC14A1 (89 genes shared exclusively out of 536 shared genes), and Exc L2 LAMP5 LTK and Exc L3−5 RORB ESR1 (83 genes shared exclusively out of 353 shared genes) (Fig 3.6C, Extended Data 4). Oligo L1-6 OPALIN and OPC L1-6 PDGFRA shared a total of 487 genes (Extended Data 4) with 68 shared exclusively (Fig 3.6C). These two cell populations also shared many DEGs with the Astro L1-6 FGFR3 SLC14A1 and Exc L2 LAMP5 LTK cell populations (~450 each) (Extended Data 4). Additionally, the two astrocyte populations, Astro L1-6 FGFR3 SLC14A1 and Astro L1-2 FGFR3 GFAP shared 378 DEGs (Extended Data 4) with 74 shared exclusively between them (Fig 3.6). 3445 genes identified as relevant to the time-series were unique to a single cell type (Fig 3.6C, Extended Data 4). The populations with the largest number of cell type-specific genes were Exc L2-3 LINC00507 FREM3 (652), Exc L2 LAMP5 LTK (638), and Astro L1-6 FGFR3 SLC14A1 (407) (Extended Data 4).

The lists of temporally regulated genes for each cell type were clustered into groups showing similar patterns of expression using a hierarchical clustering method (Supp Fig 3.20, Supp Table 3.17). The resulting patterns were then visually inspected and a subset of

similar patterns were further grouped together for GSEA (Supp Fig 3.21-3.22). The results of this analysis are summarised in Supplementary text 2. The full enrichment results are in Supp Table 3.18. While the gene lists were filtered to exclude genes expressed in fewer than 10% of nuclei in each cell type, numerous DEGs were identified that were expressed in a low percentage of nuclei (<20% of nuclei in the cell type under investigation)(Supp Table 3.17).

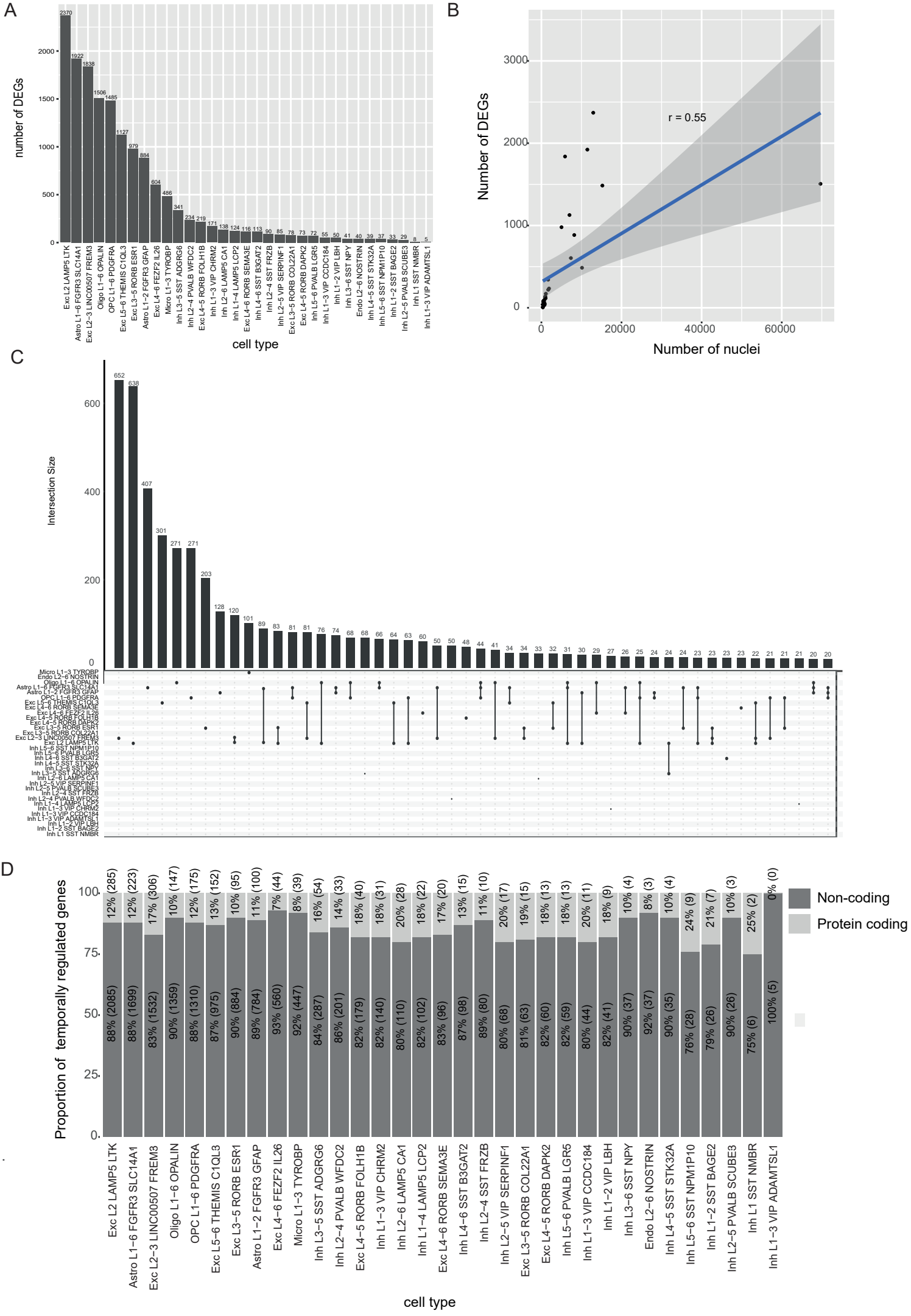

A
number of DEGs
cell type
B
Number of DEGs
Number of nuclei
r = 0.55
C
Intersection Size
D
Proportion of temporally regulated genes
cell type
Non-coding
Protein coding

Figure 3.6. Psupertime identifies temporally regulated genes in various neuronal and non-neuronal cell types. (A) Number of temporally regulated genes identified per cell type. (B) Correlation between the number of nuclei (x axis) and the number of temporally regulated genes (y axis) identified for each cell type based on Pearson's correlation. (C) Upset plots showing the top 50 intersections of temporally regulated genes unique to or overlapping the neuronal and non-neuronal cell types. The size of the intersections is indicated above the bars. The overlapping genes represent only the unique intersections between any set of cell types. (D) Proportion of temporally regulated genes per cell type that are coding or non-coding genes. The absolute number of coding and non-coding temporally regulated genes per cell type are indicated in brackets.

3.6. Combined DESeq2 and Psupertime analysis of temporally regulated genes

To narrow the DEG list down to high confidence candidates, the overlap of genes identified by DESeq2 and Psupertime for each cell type was determined, with a total of 106 genes identified by both tools across 14 cell types (Supp Table 3.19). Gene expression heatmaps were plotted for cell types with six or more consensus genes (Fig 3.7-3.8). GSEA was subsequently performed using the GO Biological processes database to identify putative functions of the genes (Fig 3.7-3.8, Supp Table 3.20). Interestingly, the consensus genes in Exc L2-3 LINC00507 FREM3 (16), Exc L4-5 RORB FOLH1B (11), Exc L3-5 RORB ESR1 (15), Oligo L1-6 OPALIN (14), and Astro L1-6 FGFR3 SLC14A1 (19), appeared to be clearly separated into two set of genes that were up- or downregulated with age (Fig 3.7-3.8). The upregulated genes in Exc L2-3 LINC00507 FREM3 included the well-described lncRNA, *LINC-PINT*, and the brain-specific angiogenesis inhibitor, *BAIAP3,* which was also upregulated in Exc L4-5 RORB FOLH1B with age (Fig 3.7A,C). Both genes not only showed an increase in the average expression level (Fig 3.7A,C; Supp Table 3.21) but also an increase in the proportion of nuclei expressing the gene with age (Fig 3.7A,C; Supp Table 3.22). Between epoch 1 and 5, the proportion of nuclei in Exc L2-3 LINC00507 FREM3 expressing *LINC-PINT* increased 2.4-fold while those expressing *BAIAP3,* increased by 6.6-fold (Supp Table 3.22). In Exc L4-5 RORB FOLH1B, the proportion of nuclei expressing *BAIAP3* increased by 9.7-fold between epoch 1 and 5 (Supp Table 3.22). GSEA revealed a role for *BAIAP3* in neurotransmitter transport and secretion (Fig 3.7A, Supp Table 3.20). Markedly, *AC002460.2* was upregulated with age in both Exc L2-3 LINC00507 FREM3 (Fig 3.7A) and Exc L2 LAMP5 LTK (Fig 3.7B) whilst *LINC01435* and *FNBP1L* were downregulated in these cell types. In addition to these genes, *MYO16* and *RASGEF1* (a gene identified as an NS-Forest marker for Exc L2-3 LINC00507 FREM3) were downregulated with age in Exc L2-3 LINC00507 FREM3 (Fig 3.7A) whereas *APOE4* was downregulated with age in Exc L2 LAMP5 LTK (Fig 3.7B). Together, *APOE4* and *FNBP1L* were associated with receptor mediated endocytosis (Fig 3.7B). In addition, *AL117329.1* (a lncRNA identified by NS-Forest as a marker for Exc L2 LAMP5 LTK) was upregulated in Exc Lamp5 LTK, Exc L4-5 RORB FOLH1B, and Exc L3-5 RORB ESR1 (Fig 3.7B-D). The proportion of nuclei expressing *AL117329.1* between epoch 1 and 5 also increased in each of these cell types (5.1-fold in Exc Lamp5 LTK, 5.9-fold in Exc L4-5 RORB FOLH1B, and 4.5-fold in Exc L3-5 RORB ESR1) (Fig 3.7B-D, Supp Table 3.22).

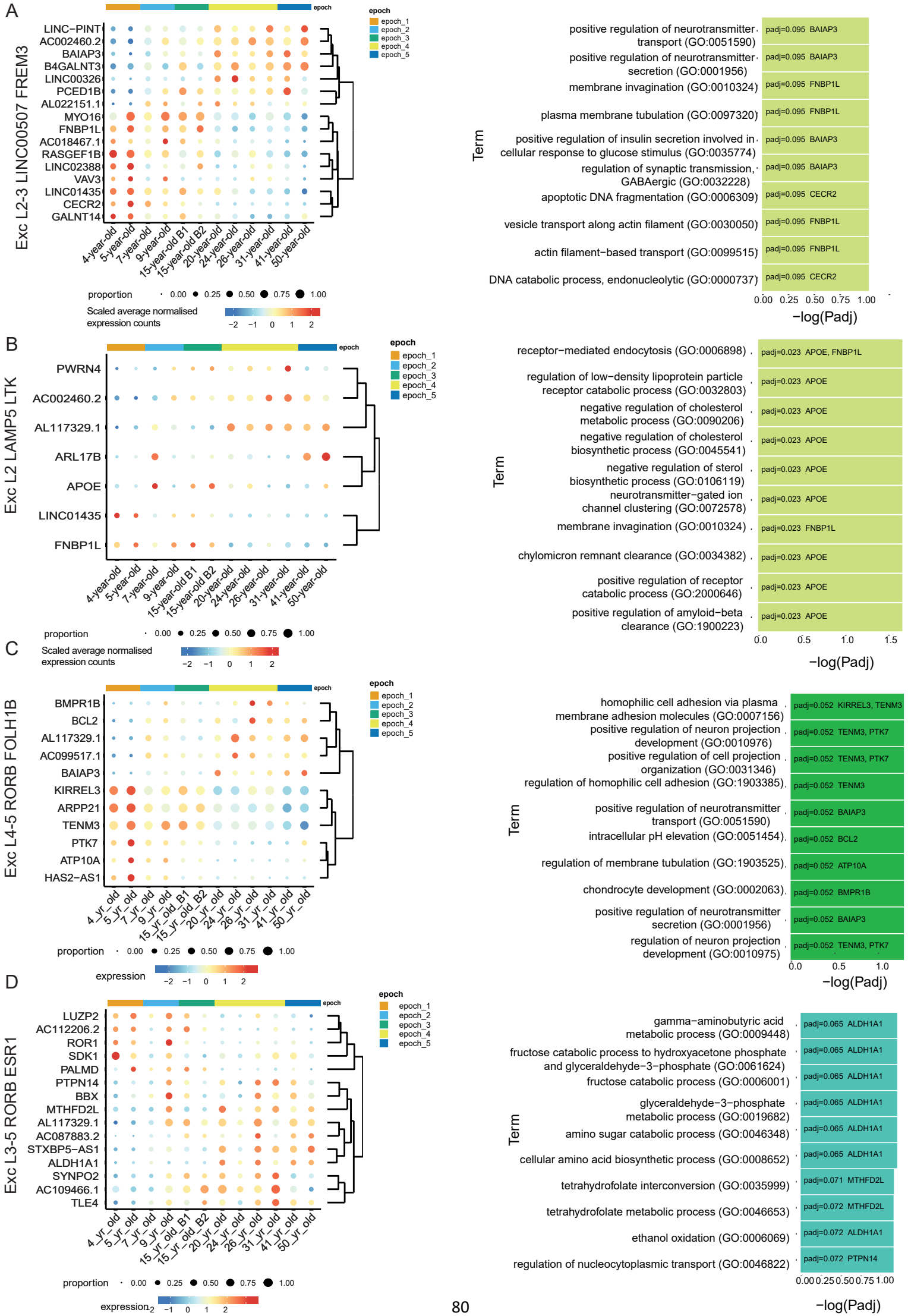

A
Exc L2-3 LINC00507 FREM3
LINC−PINT
AC002460.2
BAIAP3
B4GALNT3
LINC00326
PCED1B
AL022151.1
MYO16
FNBP1L
AC018467.1
RASGEF1B
LINC02388
VAV3
LINC01435
CECR2
GALNT14
4-year-old
5-year-old
7-year-old
9-year-old
15-year-old B1
15-year-old B2
20-year-old
24-year-old
26-year-old
31-year-old
41-year-old
50-year-old
epoch
epoch_1
epoch_2
epoch_3
epoch_4
epoch_5
proportion 0.00 0.25 0.50 0.75 1.00
Scaled average normalised expression counts −2 −1 0 1 2
Term
positive regulation of neurotransmitter transport (GO:0051590) padj=0.095 BAIAP3
positive regulation of neurotransmitter secretion (GO:0001956) padj=0.095 BAIAP3
membrane invagination (GO:0010324) padj=0.095 FNBP1L
plasma membrane tubulation (GO:0097320) padj=0.095 FNBP1L
positive regulation of insulin secretion involved in cellular response to glucose stimulus (GO:0035774) padj=0.095 BAIAP3
regulation of synaptic transmission, GABAergic (GO:0032228) padj=0.095 BAIAP3
apoptotic DNA fragmentation (GO:0006309) padj=0.095 CECR2
vesicle transport along actin filament (GO:0030050) padj=0.095 FNBP1L
actin filament−based transport (GO:0099515) padj=0.095 FNBP1L
DNA catabolic process, endonucleolytic (GO:0000737) padj=0.095 CECR2
0.00 0.25 0.50 0.75 1.00
−log(Padj)
B
Exc L2 LAMP5 LTK
PWRN4
AC002460.2
AL117329.1
ARL17B
APOE
LINC01435
FNBP1L
receptor−mediated endocytosis (GO:0006898) padj=0.023 APOE, FNBP1L
regulation of low−density lipoprotein particle receptor catabolic process (GO:0032803) padj=0.023 APOE
negative regulation of cholesterol metabolic process (GO:0090206) padj=0.023 APOE
negative regulation of cholesterol biosynthetic process (GO:0045541) padj=0.023 APOE
negative regulation of sterol biosynthetic process (GO:0106119) padj=0.023 APOE
neurotransmitter−gated ion channel clustering (GO:0072578) padj=0.023 APOE
membrane invagination (GO:0010324) padj=0.023 FNBP1L
chylomicron remnant clearance (GO:0034382) padj=0.023 APOE
positive regulation of receptor catabolic process (GO:2000646) padj=0.023 APOE
positive regulation of amyloid−beta clearance (GO:1900223) padj=0.023 APOE
0.0 0.5 1.0 1.5
−log(Padj)
C
Exc L4-5 RORB FOLH1B
BMPR1B
BCL2
AL117329.1
AC099517.1
BAIAP3
KIRREL3
ARPP21
TENM3
PTK7
ATP10A
HAS2−AS1
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
epoch
epoch_1
epoch_2
epoch_3
epoch_4
epoch_5
proportion 0.00 0.25 0.50 0.75 1.00
expression −2 −1 0 1 2
Term
homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156) padj=0.052 KIRREL3, TENM3
positive regulation of neuron projection development (GO:0010976) padj=0.052 TENM3, PTK7
positive regulation of cell projection organization (GO:0031346) padj=0.052 TENM3, PTK7
regulation of homophilic cell adhesion (GO:1903385) padj=0.052 TENM3
positive regulation of neurotransmitter transport (GO:0051590) padj=0.052 BAIAP3
intracellular pH elevation (GO:0051454) padj=0.052 BCL2
regulation of membrane tubulation (GO:1903525) padj=0.052 ATP10A
chondrocyte development (GO:0002063) padj=0.052 BMPR1B
positive regulation of neurotransmitter secretion (GO:0001956) padj=0.052 BAIAP3
regulation of neuron projection development (GO:0010975) padj=0.052 TENM3, PTK7
0.0 0.5 1.0
−log(Padj)
D
Exc L3-5 RORB ESR1
LUZP2
AC112206.2
ROR1
SDK1
PALMD
PTPN14
BBX
MTHFD2L
AL117329.1
AC087883.2
STXBP5−AS1
ALDH1A1
SYNPO2
AC109466.1
TLE4
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
epoch
epoch_1
epoch_2
epoch_3
epoch_4
epoch_5
proportion 0.00 0.25 0.50 0.75 1.00
expression −2 −1 0 1 2
Term
gamma−aminobutyric acid metabolic process (GO:0009448) padj=0.065 ALDH1A1
fructose catabolic process to hydroxyacetone phosphate and glyceraldehyde−3−phosphate (GO:0061624) padj=0.065 ALDH1A1
fructose catabolic process (GO:0006001) padj=0.065 ALDH1A1
glyceraldehyde−3−phosphate metabolic process (GO:0019682) padj=0.065 ALDH1A1
amino sugar catabolic process (GO:0046348) padj=0.065 ALDH1A1
cellular amino acid biosynthetic process (GO:0008652) padj=0.065 ALDH1A1
tetrahydrofolate interconversion (GO:0035999) padj=0.071 MTHFD2L
tetrahydrofolate metabolic process (GO:0046653) padj=0.072 MTHFD2L
ethanol oxidation (GO:0006069) padj=0.072 ALDH1A1
regulation of nucleocytoplasmic transport (GO:0046822) padj=0.072 PTPN14
0.00 0.25 0.50 0.75 1.00
−log(Padj)

Figure 3.7. High-confidence DEGs identified in excitatory neurons. (Left) Dot plot showing the expression of the shared DEGs between DESeq2 and Psupertime per sample in Exc L2-3 LINC00507 FREM3 (A), Exc L2 LAMP5 LTK (B), Exc L4-5 RORB FOLH1B (C), and Exc L3-5 RORB ESR1 (D). Data points are coloured according to the level of expression (scaled average normalised gene counts). The size of the dots indicates the proportion of nuclei expressing the gene per sample. (Right) Enrichment plots showing the top 10 enriched terms (y-axis) ranked by p-value (x-axis) which are associated with the consensus genes for Exc L2-3 LINC00507 FREM3 (A), Exc L2 LAMP5 LTK (B), Exc L4-5 RORB FOLH1B (C), and Exc L3-5 RORB ESR1 (D). GSEA plots are coloured according to the MTG taxonomy from Hodge et al. (2019).

In Exc L4-5 RORB FOLH1B, *KIRREL3*, *ARPP21, PTK7,* and *TENM3* were upregulated in paediatrics versus adults and were associated with processes such as homophilic cell adhesion via plasma membrane adhesion molecules (*KIRREL3, TENM3*) and positive regulation of neuron projection development (*TENM3, PTK7)* (Fig 3.7C). *KIRREL3*, *ARPP21,* and *TENM3* appeared to only change in their level of expression with age in Exc L4-5 RORB FOLH1B and not in the proportion of nuclei expressing them, with almost 100% of nuclei expressing each of the three gene across all epochs (Fig 3.7C, Supp Table 3.21-3.22). However, the proportion of nuclei expressing *PTK7* decreased 2.7-fold between epoch 1 and 5 (Supp Table 3.22). On the other hand, in Exc L3-5 RORB ESR1, *STXBP5−AS1, ALDH1A1*, and *SYNPO2* were downregulated in paediatrics versus adults whereas *LUZP2* and *ROR1* were upregulated in paediatrics versus adults (Fig 3.7D).

Likewise, in the Oligo L1-6 OPALIN population*, LUZP2* also showed a decreasing expression trajectory with age along with *SGCZ, NCOR2*, and *LINC01949* (Fig 3.8A). A corresponding reduction in the proportion of nuclei expressing *LUZP2* (8.3-fold) and *SGCZ* (10.4-fold) between epoch 1 and 5 was also observed (Fig 3.8A, Supp Table 3.22). In contrast to the downregulated genes*, QDPR, CDH20*, *SNCA*, *GREB1L*, and *AC008571.2* were amongst the genes upregulated with age in Oligo L1-6 OPALIN (Fig 3.8A). These were enriched for terms such as regulation of chaperone−mediated autophagy (*GFAP, SNCA)*, dopamine biosynthetic process (*SNCA*), regulation of synaptic vesicle recycling (*SNCA)*, and L-Phenylalanine catabolic process (*QDPR*) which is also required for dopamine synthesis[273] (Fig 3.8A). Curiously, the OPC L1-6 PDGFRA cluster included two genes, *HIF3A* and *AC132153.1*, which appeared to be upregulated in epoch 2, 4, and 5 whilst being downregulated in the youngest children and the adolescents (epoch 1 and 3) (Fig 3.8B). *ARL17B*, the gene shared between 27 cell types from the DESeq2 analysis (Supp Fig 3.12B), appeared to be highly expressed in epoch 5 of OPC L1-6 PDGFRA compared to all other epochs (Fig 3.8B) and this was similarly observed in Astro L1-6 FGFR3 SLC14A1 (Fig 3.8C) and Exc L2 LAMP5 LTK (Fig 3.7B).

Among the genes upregulated with age in Astro L1-6 FGFR3 SLC14A1 was *LINC00499* (Fig 3.8C) which was previously identified as a minimal marker gene for Astro L1-6 FGFR3 SLC14A1 by NS-Forest for four of the samples, including the 31-, 41-, and 50-year-old samples. Notably, there was a 6.3-fold increase in the proportion of Astro L1-6 FGFR3 SLC14A1 nuclei expressing *LINC00499* between epoch 1 and 5 (Fig 3.8C, Supp Table 3.22). In contrast to *LINC00499's* expression profile, *ELOVL5* and *TENM4* decreased in expression with age in Astro L1-6 FGFR3 SLC14A1 and were associated with terms relating to fatty acid

metabolism and oligodendrocyte differentiation, respectively (Fig 3.8C). Similarly, *ADORA2B, RELL1,* and *SLC8A1-AS1* also decreased in expression with age (Fig 3.8C). Lastly, the Micro L1-3 TYROBP included 4 genes, *RAB8B, PIK3AP1, ANKH*, and *RHEB,* which were highly expressed in the first epoch and subsequently downregulated (Fig 3.8D). GSEA suggested a role for *PIK3AP1* in toll-like receptor signalling pathways whilst *RAB8B* was associated with protein targeting to the peroxisome (Fig 3.8D). The percentage of Micro L1-3 TYROBP nuclei expressing the genes also decreased slightly between epoch 1 and 5, by 1.6-fold for *RAB8B,* 2.3-fold for *PIK3AP1,* 2.2-fold for *ANKH,* and 2.1 fold for *RHEB* (Fig 3.8D, Supp Table 3.22). On the other hand, the lncRNA, *LINC02232*, showed an opposing trend in the proportion of nuclei expressing it (11.3-fold increase) and in the average expression level in Micro L1-3 TYROBP (Fig 3.8D, Supp Table 3.21-3.22).

Taken together, the consensus genes identified between the Psupertime and DESeq2 analyses represent high-confidence temporally regulated genes which were either shared across multiple different cell types or specific to certain cell types. These may have important functions in controlling postnatal maturational processes in the brain.

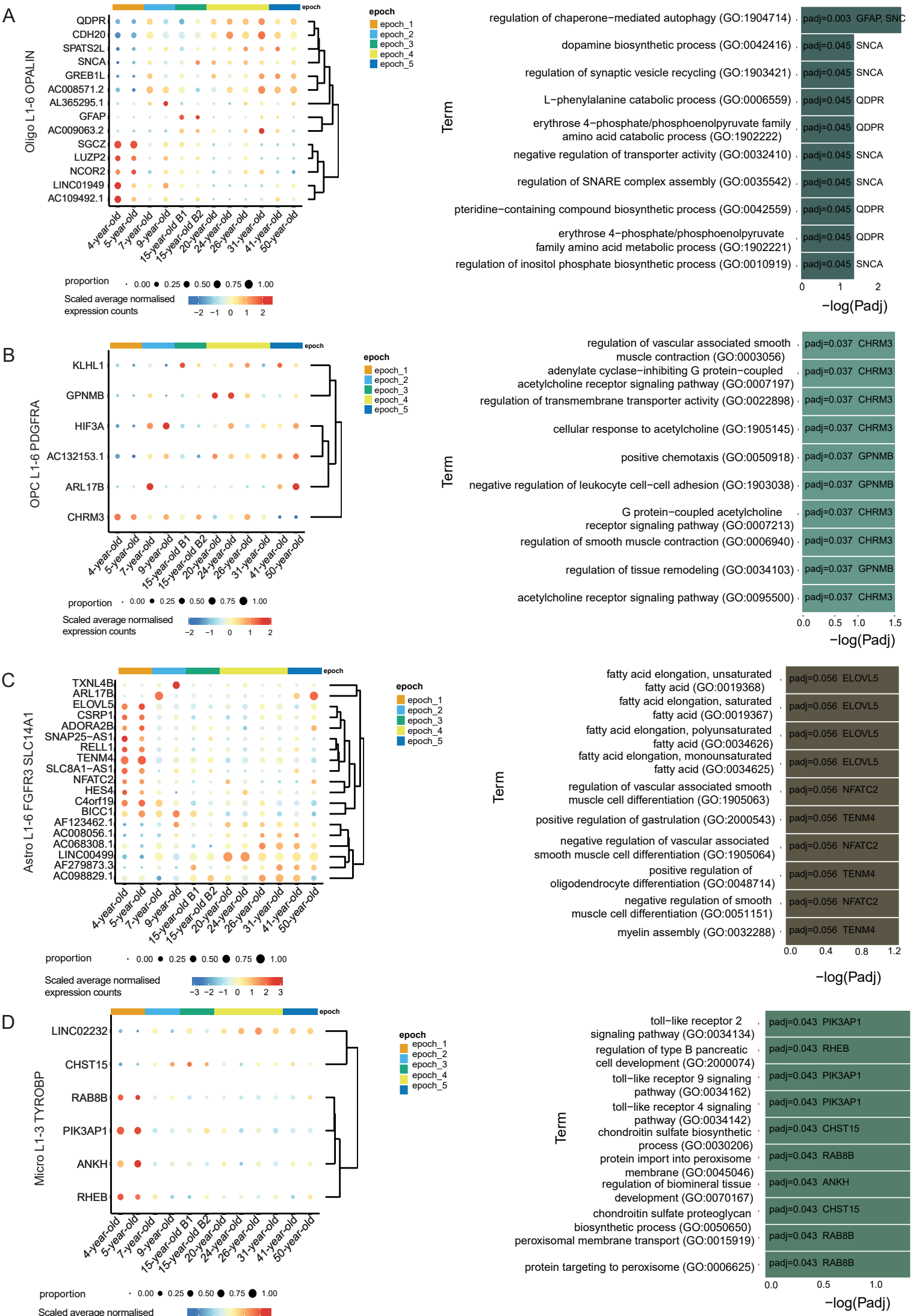

A
Oligo L1-6 OPALIN
QDPR
CDH20
SPATS2L
SNCA
GREB1L
AC008571.2
AL365295.1
GFAP
AC009063.2
SGCZ
LUZP2
NCOR2
LINC01949
AC109492.1
4-year-old
5-year-old
7-year-old
9-year-old
15-year-old B1
15-year-old B2
20-year-old
24-year-old
26-year-old
31-year-old
41-year-old
50-year-old
epoch
epoch_1
epoch_2
epoch_3
epoch_4
epoch_5
proportion
0.00 0.25 0.50 0.75 1.00
Scaled average normalised expression counts
−2 −1 0 1 2
Term
regulation of chaperone−mediated autophagy (GO:1904714)
dopamine biosynthetic process (GO:0042416)
regulation of synaptic vesicle recycling (GO:1903421)
L−phenylalanine catabolic process (GO:0006559)
erythrose 4−phosphate/phosphoenolpyruvate family amino acid catabolic process (GO:1902222)
negative regulation of transporter activity (GO:0032410)
regulation of SNARE complex assembly (GO:0035542)
pteridine−containing compound biosynthetic process (GO:0042559)
erythrose 4−phosphate/phosphoenolpyruvate family amino acid metabolic process (GO:1902221)
regulation of inositol phosphate biosynthetic process (GO:0010919)
padj=0.003 GFAP, SNC
padj=0.045 SNCA
padj=0.045 SNCA
padj=0.045 QDPR
padj=0.045 QDPR
padj=0.045 SNCA
padj=0.045 SNCA
padj=0.045 QDPR
padj=0.045 QDPR
padj=0.045 SNCA
0 1 2
−log(Padj)
B
OPC L1-6 PDGFRA
KLHL1
GPNMB
HIF3A
AC132153.1
ARL17B
CHRM3
regulation of vascular associated smooth muscle contraction (GO:0003056)
adenylate cyclase−inhibiting G protein−coupled acetylcholine receptor signaling pathway (GO:0007197)
regulation of transmembrane transporter activity (GO:0022898)
cellular response to acetylcholine (GO:1905145)
positive chemotaxis (GO:0050918)
negative regulation of leukocyte cell−cell adhesion (GO:1903038)
G protein−coupled acetylcholine receptor signaling pathway (GO:0007213)
regulation of smooth muscle contraction (GO:0006940)
regulation of tissue remodeling (GO:0034103)
acetylcholine receptor signaling pathway (GO:0095500)
padj=0.037 CHRM3
padj=0.037 CHRM3
padj=0.037 CHRM3
padj=0.037 CHRM3
padj=0.037 GPNMB
padj=0.037 GPNMB
padj=0.037 CHRM3
padj=0.037 CHRM3
padj=0.037 GPNMB
padj=0.037 CHRM3
0.0 0.5 1.0 1.5
−log(Padj)
C
Astro L1-6 FGFR3 SLC14A1
TXNL4B
ARL17B
ELOVL5
CSRP1
ADORA2B
SNAP25−AS1
RELL1
TENM4
SLC8A1−AS1
NFATC2
HES4
C4orf19
BICC1
AF123462.1
AC008056.1
AC068308.1
LINC00499
AF279873.3
AC098829.1
−3 −2 −1 0 1 2 3
fatty acid elongation, unsaturated fatty acid (GO:0019368)
fatty acid elongation, saturated fatty acid (GO:0019367)
fatty acid elongation, polyunsaturated fatty acid (GO:0034626)
fatty acid elongation, monounsaturated fatty acid (GO:0034625)
regulation of vascular associated smooth muscle cell differentiation (GO:1905063)
positive regulation of gastrulation (GO:2000543)
negative regulation of vascular associated smooth muscle cell differentiation (GO:1905064)
positive regulation of oligodendrocyte differentiation (GO:0048714)
negative regulation of smooth muscle cell differentiation (GO:0051151)
myelin assembly (GO:0032288)
padj=0.056 ELOVL5
padj=0.056 ELOVL5
padj=0.056 ELOVL5
padj=0.056 ELOVL5
padj=0.056 NFATC2
padj=0.056 TENM4
padj=0.056 NFATC2
padj=0.056 TENM4
padj=0.056 NFATC2
padj=0.056 TENM4
0.0 0.4 0.8 1.2
−log(Padj)
D
Micro L1-3 TYROBP
LINC02232
CHST15
RAB8B
PIK3AP1
ANKH
RHEB
toll−like receptor 2 signaling pathway (GO:0034134)
regulation of type B pancreatic cell development (GO:2000074)
toll−like receptor 9 signaling pathway (GO:0034162)
toll−like receptor 4 signaling pathway (GO:0034142)
chondroitin sulfate biosynthetic process (GO:0030206)
protein import into peroxisome membrane (GO:0045046)
regulation of biomineral tissue development (GO:0070167)
chondroitin sulfate proteoglycan biosynthetic process (GO:0050650)
peroxisomal membrane transport (GO:0015919)
protein targeting to peroxisome (GO:0006625)
padj=0.043 PIK3AP1
padj=0.043 RHEB
padj=0.043 PIK3AP1
padj=0.043 PIK3AP1
padj=0.043 CHST15
padj=0.043 RAB8B
padj=0.043 ANKH
padj=0.043 CHST15
padj=0.043 RAB8B
padj=0.043 RAB8B
0.0 0.5 1.0
−log(Padj)

Figure 3.8. High-confidence DEGs identified in glial cells. (Left) Dot plot showing the expression of the shared DEGs between DESeq2 and Psupertime per sample in Oligo L1-6 OPALIN (A), OPC L1-6 PDGFRA (B), Astro L1-6 FGFR3 SLC14A1 (C), and Micro L1-3 TYROBP (D). Data points are coloured according to the level of expression (scaled average normalised gene counts). The size of the dots indicates the proportion of nuclei expressing the gene per sample. (Right) Enrichment plots showing the top 10 enriched terms (y-axis) ranked by p-value (x-axis) which are associated with the consensus genes for Oligo L1-6 OPALIN (A), OPC L1-6 PDGFRA (B), Astro L1-6 FGFR3 SLC14A1 (C), and Micro L1-3 TYROBP (D). GSEA plots are coloured according to the MTG taxonomy from Hodge et al. (2019).

3.7. Differential gene expression analysis between paediatric and adult samples within specific cell types using IDEAS

In addition to the DESeq2 LRT method and Psupertime, an independent differential expression analysis method known as Individual level Differential Expression Analysis for Single cells (IDEAS)[222] was applied to identify DEGs between the paediatric and adult samples (Fig 3.9). This analysis was performed in a pairwise fashion instead of a multi-group comparison between the different epochs as the tool is currently only designed to compare two groups. A total of 12 896 DEGs were identified overall and, despite having fewer nuclei than many other populations, several inhibitory neuron subtypes had the largest number of DEGs including Inh L5−6 SST MIR548F2 (497), Inh L3−6 SST NPY (476), and Inh L4−6 SST GXYLT2 (472) (Fig 3.9A). In fact, correlating the number of DEGs identified by IDEAS to the number of nuclei per population revealed a negative correlation between these parameters (r = -0.347) (Fig 3.9B). This correlation score remained comparable even after removing the outlying oligodendrocyte population (r=-0.329) (Supp Fig 3.12D). An analysis of the overlapping DEGs between cell types revealed numerous DEGs to be specific to a single cell type (Fig 3.9C). Altogether, 3527 DEGs were identified as cell type-specific with the greatest number of unique genes observed in Inh L4-6 SST GXYLT2 (227), Inh L5-6 SST MIR548F2(192), and Exc L4-5 RORB DAPK2 (157)(Fig 3.9C, Extended Data 5). The populations sharing the greatest number of DEGs appeared to be subtypes of excitatory neurons including Exc L2 LAMP5 LTK and Exc L2-3 LINC00507 FREM3 (43 shared DEGs with 4 shared exclusively), Exc L2-3 LINC00507 FREM3 and Exc L4-5 RORB FOLH1B (43 shared DEGs with 5 shared exclusively), Exc L2 LAMP5 LTK and Exc L3-5 RORB ESR1 (41 shared DEGs with 5 shared exclusively), as well as Exc L2-3 LINC00507 FREM3 and Exc L3-5 RORB ESR1 (40 shared DEGs with 6 shared exclusively) (Extended Data 5). Notably, more than 68% of the DEGs in any population were protein-coding (Fig 3.9D).

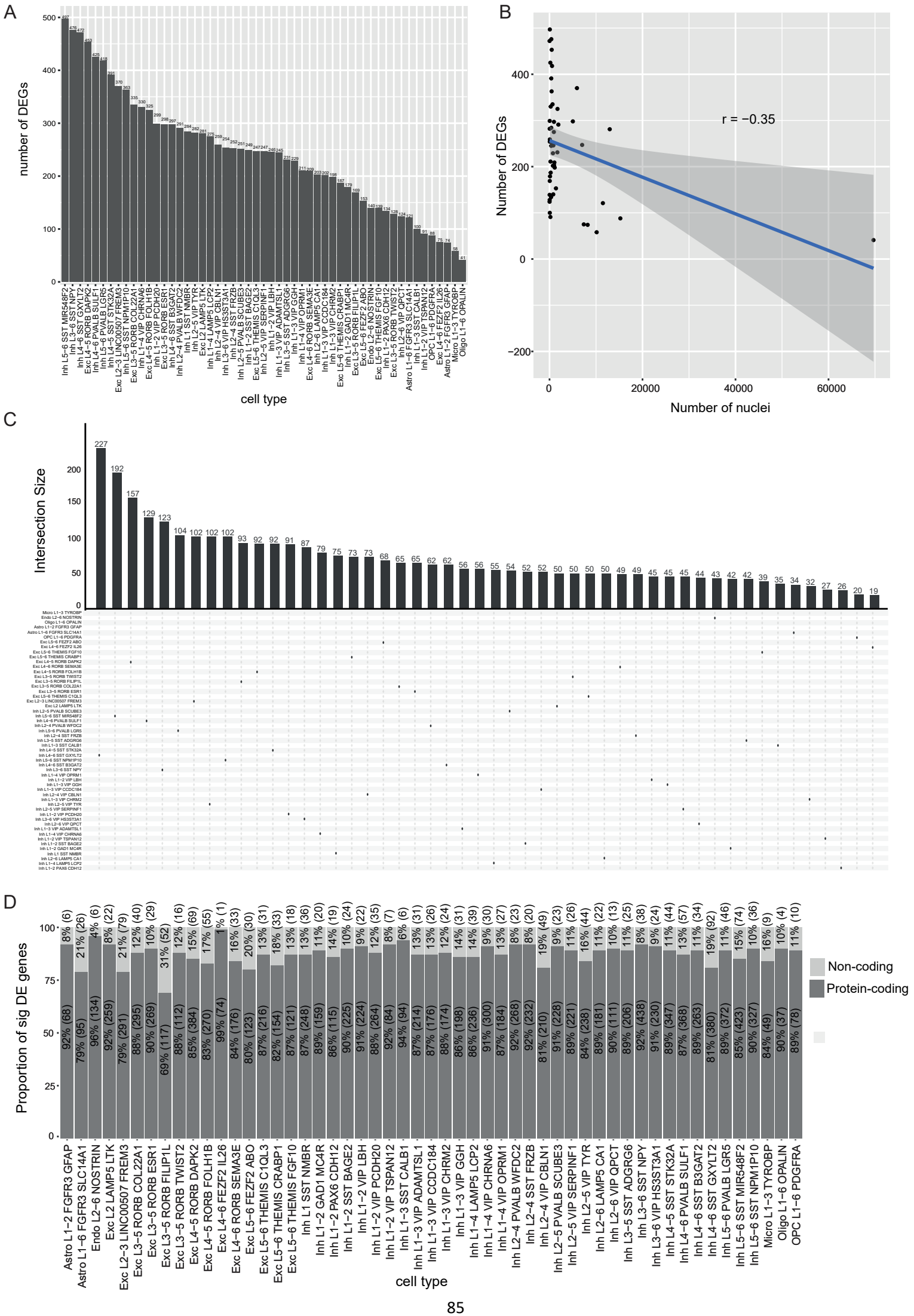

A
number of DEGs
cell type
B
Number of DEGs
r = −0.35
Number of nuclei
C
Intersection Size
D
Proportion of sig DE genes
cell type
Non-coding
Protein-coding

Figure 3.9. IDEAS identifies DEGs in neuronal and non-neuronal cell types. (A) Number of DEGs identified per cell type. (B) Correlation between the number of nuclei and the number of DEGs identified for each cell type based on Pearson's correlation. (C) Upset plots showing the top 50 sets of DEGs, which were unique the neuronal and non-neuronal cell types (i.e. no intersections between cell types were present for the top 50 DEG sets). (D) Proportion of DEGs per cell type that are coding or non-coding genes.

Heatmaps of the significant DEGs showed largely uniform patterns comprising a set of up and downregulated genes between the two age groups for each cell type (Fig 3.10-3.11, Supp Fig 3.23-3.25, Supp Fig 3.29, Supp Table 3.23). There were several cell types which did not have nuclei for each of the samples, but significant genes were identified using a subset of the samples which did have nuclei (Supp Fig 3.29, Supp Table 3.6). Inh L2-4 VIP SPAG17 had too few nuclei to perform the analysis. GSEA was performed with Enrichr[231] on the lists of up and downregulated genes for each cell type. A subset of cell types were selected for further investigation including all 6 non-neuronal sub-types, as well as the top 6 excitatory sub-types and top 6 inhibitory sub-types based on the total number of nuclei per cell type. Of these cell types, the results of 8 cell types of interest are described in the main text (Fig 3.10-3.14) while the rest of the results are described in the supplementary data (Supp Fig 3.23-3.28, Supplementary text 3). For the enrichment analysis, the top 5 terms per cell type per database for the up- and downregulated genes is summarised (Fig 3.12-3.14). The full outputs of the IDEAS analysis and associated enrichment analysis are in the supplementary data (Supp Fig 3.29, Supp Table 3.23-3.29).

In the Astro L1-2 FGFR3 GFAP population, the set of genes that was upregulated with age (27) was associated with GO biological processes such as regulation of membrane tubulation (*WASL*), regulation of extracellular exosome assembly (*SDC4*), protein-containing complex subunit organization (*CAPZA1, WASL*), and negative regulation of lymphocyte migration (WASL) (Fig 3.10A, Fig 3.12A, Supp Table 3.24). Additionally, these genes were associated with diseases such as Distal Hereditary Motor Neuropathy, Charcot-Marie-Tooth disease, and Spinal muscular atrophy (*HSPB8*) (Fig 3.10A, Fig 3.13A, Supp Table 3.26). Genes downregulated with age in Astro L1-2 FGFR3 GFAP (47) were enriched for GO biological processes including ephrin receptor signalling (*FYN, KALRN, DNM1*), phosphorylation (*NIM1K, DGKB, STK38L, FYN, HK1*), and postsynaptic neurotransmitter receptor internalization (*DNM1*) (Fig 3.10A, Fig 3.12B, Supp Table 3.25). Additionally, these genes were associated with neurological conditions such as Schizophrenia (*MAP6, SLC6A1*) and temporal lobe epilepsy (*GRIK5, FYN, SLC6A1*) (Fig 3.10A, Fig 3.13B, Supp Table 3.27).

In Endo L2-6 NOSTRIN, the genes that were more highly expressed in the paediatric samples than the adult samples (41) were enriched for GO biological processes such as contractile actin filament bundle assembly (*FAM171A1, ITGB5*), calcium ion transmembrane transport (*SLC24A2, CACNA1C, ATP2B1*), and regulation of transmembrane transport (*KCNJ6, PRKCB*) (Fig 3.10B, Fig 3.12B, Supp Table 3.25). Moreover, these genes were enriched for DisGeNET terms including high density lipoprotein measurement (*ABCC4, CACNA1C, ATP2B1, ATG7, UBE2L3*), mental depression (*KCNJ6, SUCLA2, CHL1, PRKCB, ACSL4, CACNA1C*), and Parkinson's disease (*RB1, USP15, KCNJ6, SUCLA2, VDAC1, EIF2S2, ATG7, UBE2L3*) (Fig 3.10B, Fig 3.13B, Supp Table 3.27). Genes upregulated with age in Endo L2-6 NOSTRIN (99) were

associated with negative regulation of TOR signalling (*DEPTOR, TSC1, GAS6, FNIP1, FBXO9*), regulation of organelle assembly (*DYNC2LI1, PIKFYVE, RABGAP1, TSG101*), glycine transport (*SLC7A8, SLC38A5*), positive regulation of response to cytokine stimulus (IFIH1, GAS6) (Fig 3.10B, Fig 3.12A, Supp Table 3.24). These genes were also enriched for DisGeNET terms such as Neurofibromatosis (*VCP, CREB1, PRKAR1A, DPYSL2, SDHC, TSC1, PGR*) and mental deficiency (*KDM5C, ASAH1, NR2F1, SDHC, TSC1, CPLANE1, THOC2, CTCF, IFIH1, PRPF6, ARL2BP, PRKAR1A, RARS, DNAJC21*) (Fig 3.10B, Fig 3.13A, Supp Table 3.26).

The set of genes upregulated with age in Oligo L1-6 OPALIN (31) was associated with negative regulation of immune effector process (*CD55, APPL1*), neuron projection morphogenesis (*SHTN1, ALCAM, LRP2*), spliceosomal complex assembly (*SCAF11, RBM5*), and various terms relating to cell adhesion (*ALCAM, CADM1, CDH20, PKN2*) (Fig 3.10C, Fig 3.12A, Supp Table 3.24). In contrast, the genes downregulated with age in Oligo L1-6 OPALIN (10) were enriched for processes such as positive regulation of cell cycle (*DYNC1H1, PKP4*) and positive regulation of hydrolase activity (*PKP4, EVI5L*) (Fig 3.10C, Fig 3.12B, Supp Table 3.25) as well as diseases such as Spinal Muscular Atrophies of Childhood (*DYNC1H1*) (Fig 3.10C, Fig 3.13B, , Supp Table 3.27). In the OPC L1-6 PDGFRA population, terms associated with genes that were upregulated with age (42) include autophagy (*ATG13, ATG2B*), supramolecular fiber organization (*COL5A3, CRIPT, NF1, MYO5A, SHROOM4*), spliceosomal complex assembly (*GCFC2, RBM5*), and cognition (*NF1, SHROOM4*) (Fig 3.10D, Fig 3.12A, Supp Table 3.24). The genes upregulated in the paediatric samples compared to the adult samples (46) were associated with GO biological processes such as axonal transport (*DYNC1H1, KIF1A, SYBU*), cell junction organization (*DYNC1H1, KIF1A, SYBU*), vesicle transport along microtubule (*DYNC1H1, KIF1A*), calcium ion transport (*DYNC1H1, KIF1A*), and synapse organization (*SLC8A3, UNC13C, SLC6A1*) (Fig 3.10D, Fig 3.12B, Supp Table 3.25). Furthermore, they were enriched for neurological conditions such as common migraine (*SLC8A3, UNC13C, SLC6A1*), epilepsy (*GRIA1, DYNC1H1, SLC35A3, NPRL3, CELF4, GOPC, ACO2, SLC6A1, ALDH7A1, CNN3*), cortical dysplasia (*DYNC1H1, LINGO1, NPRL3*), movement disorders (*DYNC1H1, LINGO1, NPRL3*), and neurodegenerative diseases (*DYNC1H1, LINGO1, NBR1, KIF1A, ACO2, SNCAIP, SYBU*) (Fig 3.10D, Fig 3.13B, Supp Table 3.27).

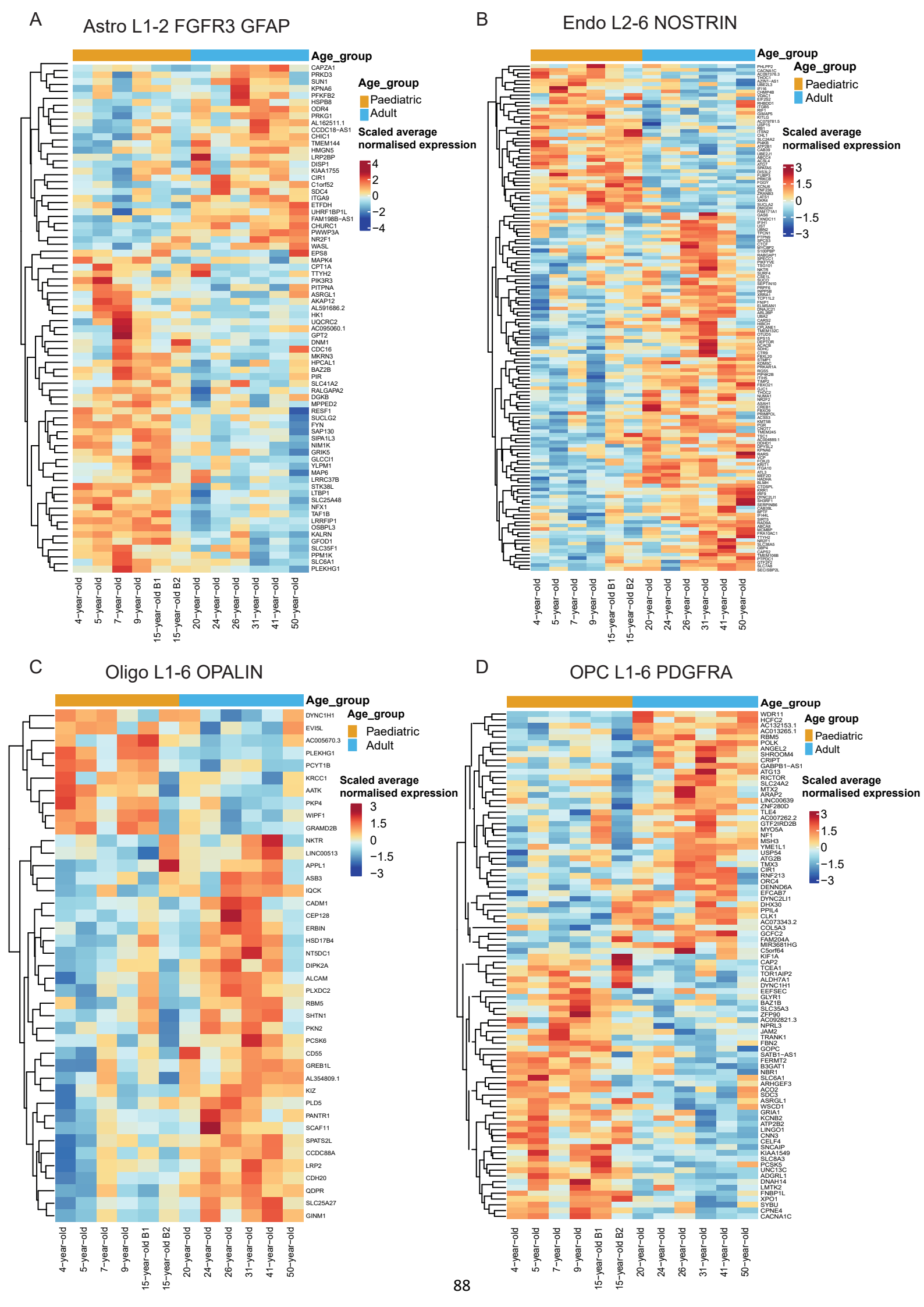

A
Astro L1-2 FGFR3 GFAP
Age_group
Paediatric
Adult
Scaled average normalised expression
4
2
0
−2
−4
CAPZA1
PRKD3
SUN1
KPNA6
PFKFB2
HSPB8
ODR4
PRKG1
AL162511.1
CCDC18−AS1
CHIC1
TMEM144
HMGN5
LRP2BP
DISP1
KIAA1755
CIR1
C1orf52
SDC4
ITGA9
ETFDH
UHRF1BP1L
FAM198B−AS1
CHURC1
PWWP3A
NR2F1
WASL
EPS8
MAPK4
CPT1A
TTYH2
PIK3R3
PITPNA
ASRGL1
AKAP12
AL591686.2
HK1
UQCRC2
AC095060.1
GPT2
DNM1
CDC16
MKRN3
HPCAL1
BAZ2B
PIR
SLC41A2
RALGAPA2
DGKB
MPPED2
RESF1
SUCLG2
FYN
SAP130
SIPA1L3
NIM1K
GRIK5
GLCCI1
YLPM1
MAP6
LRRC37B
STK38L
LTBP1
SLC25A48
NFX1
TAF1B
LRRFIP1
OSBPL3
KALRN
GFOD1
SLC35F1
PPM1K
SLC6A1
PLEKHG1
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old
B
Endo L2-6 NOSTRIN
Age_group
Paediatric
Adult
Scaled average normalised expression
3
1.5
0
−1.5
−3
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old
C
Oligo L1-6 OPALIN
Age_group
Paediatric
Adult
Scaled average normalised expression
3
1.5
0
−1.5
−3
DYNC1H1
EVI5L
AC005670.3
PLEKHG1
PCYT1B
KRCC1
AATK
PKP4
WIPF1
GRAMD2B
NKTR
LINC00513
APPL1
ASB3
IQCK
CADM1
CEP128
ERBIN
HSD17B4
NT5DC1
DIPK2A
ALCAM
PLXDC2
RBM5
SHTN1
PKN2
PCSK6
CD55
GREB1L
AL354809.1
KIZ
PLD5
PANTR1
SCAF11
SPATS2L
CCDC88A
LRP2
CDH20
QDPR
SLC25A27
GINM1
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old
D
OPC L1-6 PDGFRA
Age_group
Age group
Paediatric
Adult
Scaled average normalised expression
3
1.5
0
−1.5
−3
WDR11
HCFC2
AC132153.1
AC013265.1
RBM5
POLK
ANGEL2
SHROOM4
CRIPT
GABPB1−AS1
ATG13
RICTOR
SLC24A2
MTX2
ARAP2
LINC00639
ZNF280D
TLE4
AC007262.2
GTF2IRD2B
MYO5A
NF1
MSH3
YME1L1
USP54
ATG2B
TMX3
CIR1
RNF213
ORC4
DENND6A
EFCAB7
DYNC2LI1
DHX30
PPIL4
CLK1
AC073343.2
COL5A3
GCFC2
FAM204A
MIR3681HG
C5orf64
KIF1A
CAP2
TCEA1
TOR1AIP2
ALDH7A1
DYNC1H1
EEFSEC
GLYR1
BAZ1B
SLC35A3
ZFP90
AC092821.3
NPRL3
JAM2
TRANK1
FBN2
GOPC
SATB1−AS1
FERMT2
B3GAT1
NBR1
SLC6A1
ARHGEF3
ACO2
SDC3
ASRGL1
WSCD1
GRIA1
KCNB2
ATP2B2
LINGO1
CNN3
CELF4
SNCAIP
KIAA1549
SLC8A3
PCSK5
UNC13C
ADGRL1
DNAH14
LMTK2
FNBP1L
XPO1
SYBU
CPNE4
CACNA1C
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old

Figure 3.10. IDEAS reveals sets of genes that are up- and downregulated with age in several non-neuronal cell types. Heatmap plots showing the change in the level of expression with age of DEGs for Astro L1-2 FGFR3 GFAP (A), Endo L2-6 NOSTRIN (B), Oligo L1-6 OPALIN (C), and OPC L1-6 PDGFRA (D). Expression level is the scaled average normalised gene counts per sample within each cell type. X axes represents samples. Y axes represents DEGs.

In Exc L2 LAMP5 LTK, genes upregulated in the adult versus paediatric samples (155) were enriched for GO biological terms such as protein localization to membrane (*CACNB2, DLG1, EPB41L3, MAP7, ANK2, FAM126B, ZDHHC2, ASB3*), regulation of cation channel activity (*CACNB2, DLG1, FGF14, PDE4D, ANK2*), and chemical synaptic transmission (*GABRA2, CACNB2, DLG1, GABRB1, HTR1E, PRKAR2B, AMPH, HTR4, SYN3*) (Fig 3.11A, Fig3.12A, Supp Table 3.24). On the other hand, GO biological terms associated with genes that were downregulated in the adult versus paediatric samples (126) included regulation of telomere maintenance via telomere lengthening (*PARP1, TNKS2, DHX36*) and transport across blood brain barrier (*SLC27A1, SLC24A3, SLC4A3, SLC7A1, SLC8A2*) (Fig 3.11A, Fig 3.12B, Supp Table 3.25). Interestingly, the downregulated genes were also associated with Alzheimer's disease (*MEF2C, PARP1, BRI3, LUZP2, KIDINS220, GOLM1, TXNRD1, PTGER3, PRRC2C, CRMP1, NRG1, SNCAIP, SREBF2, MYO16, YY1, POLB, BACE1, SUCLA2, NDUFAF6, DLG3, DNAJC5, DPYSL3, MADD, ANXA6*) as well as mental retardation (*YY1, DYNC1H1, MEF2C, PARP1, FOXG1*) (Fig 3.11A, Fig 3.13B, Supp Table 3.27). Lastly, several of these genes have previously been shown to be downregulated in the brain with age between 20-year-olds versus 50-year-olds (*ROBO2, SVOP, ZNF804B, NRG1, SLC8A2*), 60-year-olds (*ROBO2, FSTL5, SVOP, ZNF804B, NRG1*), and 70-year-olds (*SVOP, PTGER3, ZNF804B, NRG1*) (Fig 3.11A, Fig 3.14B, Supp Table 3.29).

Similarly to Exc L2 LAMP5 LTK, the set of genes downregulated in Exc L3-5 RORB ESR1 (119) also included genes from the GTEx database that are downregulated in the brain between 20 year olds versus 60 year olds (*ROBO2, SVOP, PLPPR4, NREP, SLC8A2, RNF165*) and 70 year olds (*ROBO2, SVOP, NREP*) (Fig 3.11B, Fig 3.14B, Supp Table 3.29). Additionally, these genes were associated with GO biological processes such as telomere maintenance (*RIF1, PARP1, XRCC5, TERF2IP*) and neuron migration (*MEF2C, DCX, FGF13, PAFAH1B1*) (Fig 3.11B, Fig 3.12B, Supp Table 3.25). Moreover, they were associated with various neurological conditions such as cerebrovascular disorders (*MEF2C, JPH3, DCX, CYP46A1, SORL1, PAFAH1B1*), subcortical band heterotopia (*DCX, PAFAH1B1*), neurodevelopmental disorders (*KMT2D, MEF2C, USP7, PTCHD1, BCL11A, PTPN11*), Subependymal Giant Cell Astrocytoma (*KIAA1549, DCX, PTPN11*), and childhood neuroblastoma (*PARP1, DCX, CRMP1, PTPN11*) (Fig 3.11B, Fig 3.13B, Supp Table 3.27). The set of genes upregulated with age in Exc L3-5 RORB ESR1 (179) were enriched for GO Biological processes such as regulation of alternative mRNA splicing via spliceosome (*MBNL1, MBNL2, NOVA1, TRA2B*), regulation of ryanodine−sensitive calcium release channel activity (*CAMK2D, PDE4D, PKD2*), and dendritic spine maintenance (*MTMR2, IGF1R*) (Fig 3.11B, Fig 3.12A, Supp Table 3.24) whilst also being enriched for genes associated with Huntington disease-like2 (*RIMS2, MBNL1, MBNL2*) and hippocampal atrophy (*LHFPL6, PRUNE2*) (Fig 3.11B, Fig 3.13A, Supp Table 3.26).

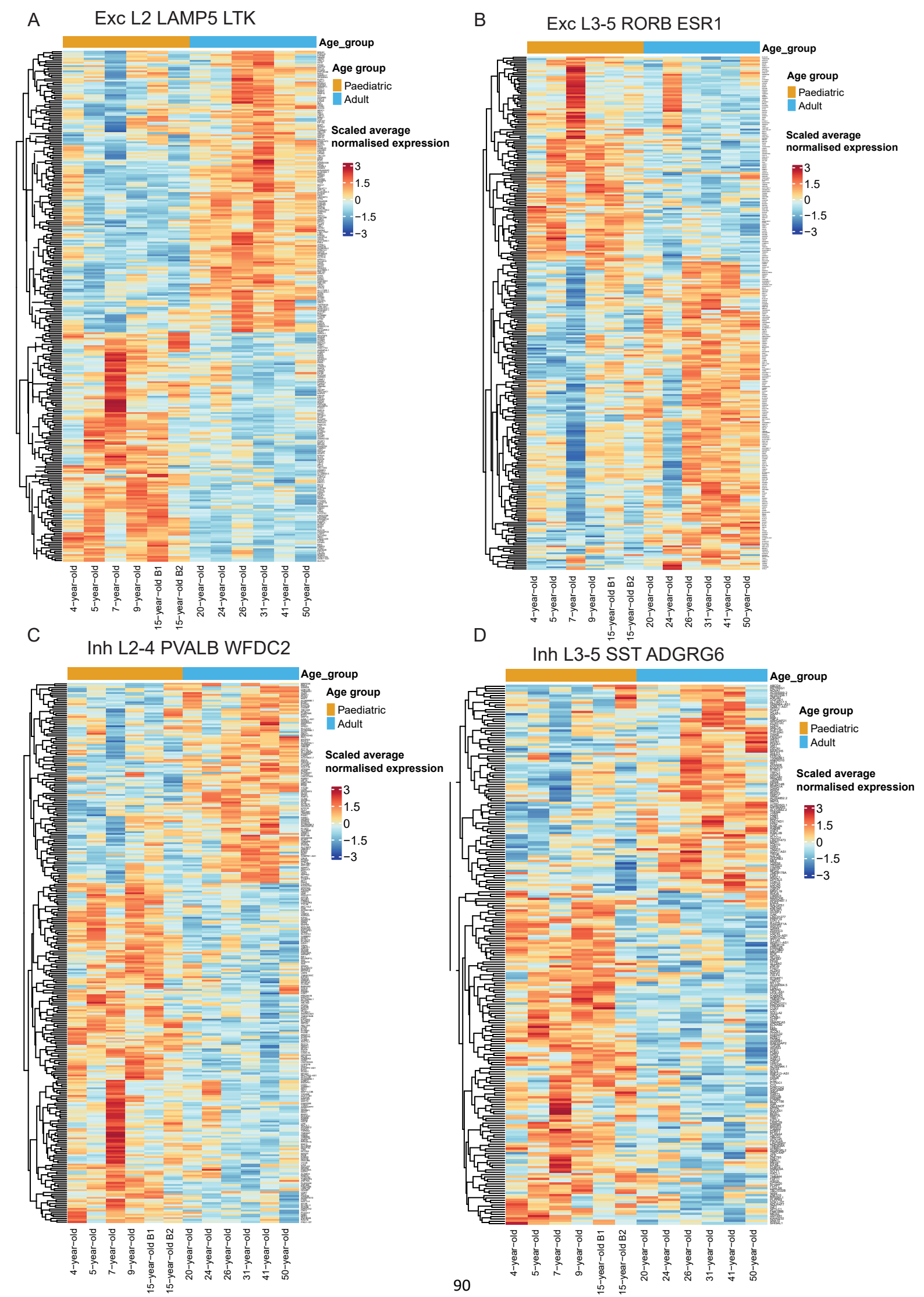

A
Exc L2 LAMP5 LTK
B
Exc L3-5 RORB ESR1
C
Inh L2-4 PVALB WFDC2
D
Inh L3-5 SST ADGRG6
Age_group
Age group
Paediatric
Adult
Scaled average normalised expression
3
1.5
0
−1.5
−3
4−year−old
5−year−old
7−year−old
9−year−old
15−year−old B1
15−year−old B2
20−year−old
24−year−old
26−year−old
31−year−old
41−year−old
50−year−old

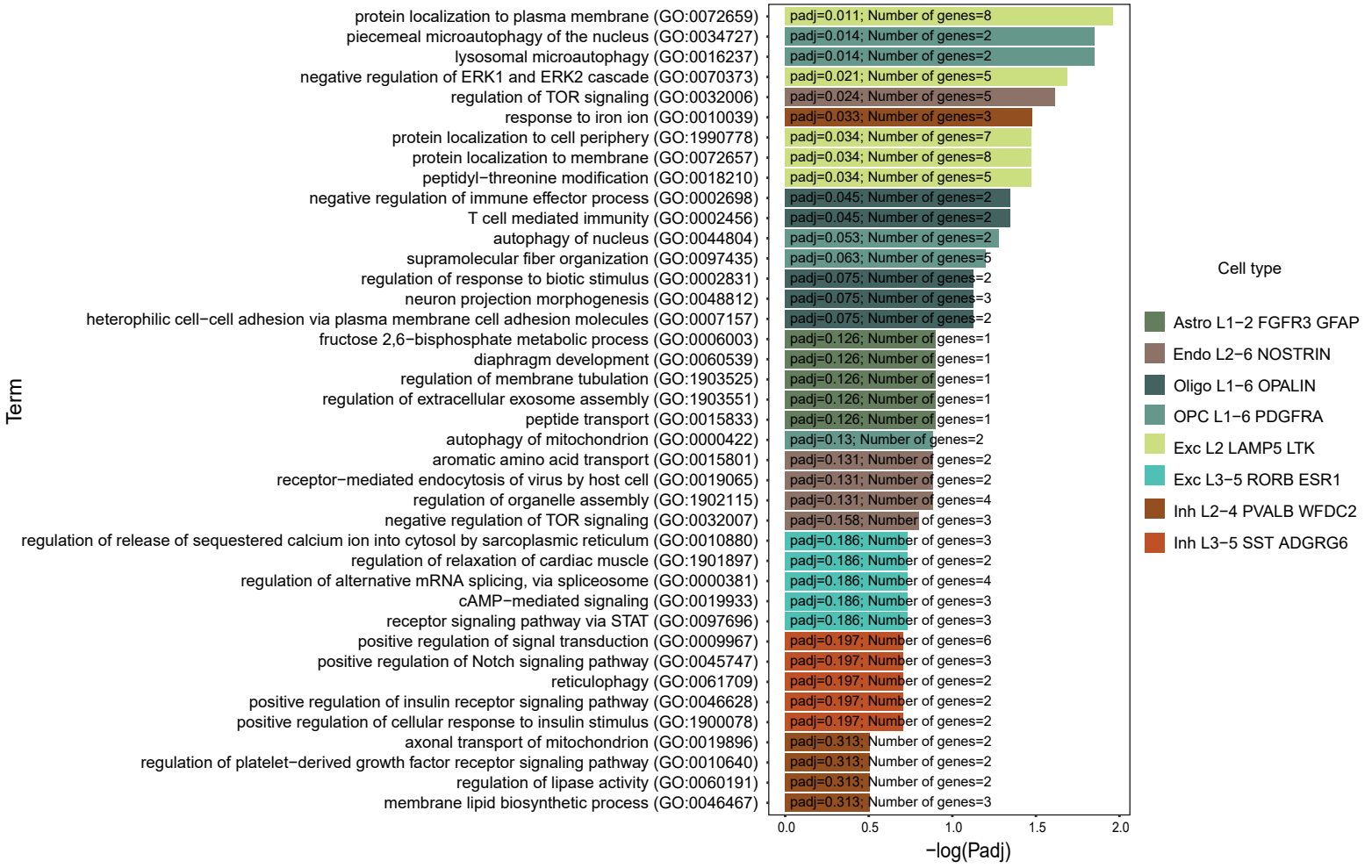
Figure 3.11. IDEAS reveals sets of genes that are up- and downregulated with age in several neuronal cell types. Heatmap plots showing the change in the level of expression with age of DEGs for Exc L2 LAMP5 LTK (A), Exc L3-5 RORB ESR1 (B), Inh L2-4 PVALB WFDC2 (C), and Inh L3-5 SST ADGRG6 (D). Expression level is the scaled average normalised gene counts per sample within each cell type. The samples were grouped into either the paediatric or adult age group. X axes represents samples. Y axes represents DEGs.

For the Inh L2-4 PVALB WFDC2 population, genes increasing in their level of expression as the brain matures (108) were associated with processes such as response to iron ion (*ACO1, HIF1A, SNCA*) and membrane lipid biosynthetic process (*SAMD8, PPM1L, PLPP1*) (Fig 3.11C, Fig 3.12B, Supp Table 3.24) whereas genes showing an opposing trajectory were associated with glycerophospholipid biosynthetic process (*CDS1, PGS1, SLC27A1, PCYT1A, LPIN2, PIK3C2B, AGPAT4, PLEKHA8*), clathrin−dependent endocytosis (*GAK, CLTC, AP2A1*) (Fig 3.11C, Fig 3.12A, Supp Table 3.25). On the other hand, the downregulated genes in Inh L2-4 PVALB WFDC2 (183) were enriched for numerous genes from the GTEx Aging Signatures database which are downregulated in the brain with age (*GABRA1, GABBR2, RIMS4, GAP43, NWD2, LRRC38, ZNF804B, NMNAT2, SLC8A2*) (Fig 3.11C, Fig 3.14, Supp Table 3.29).

Likewise, genes that were downregulated with age in Inh L3-5 SST ADGRG6 (143) were also found to be downregulated in the brain with age according to the GTEx database (*MYT1L, WSCD2, PITPNM3, PLPPR4, CAMK4, SHISAL1, MEIS3, SLC8A2*) (Fig 3.11D, Fig 3.14B, Supp Table 3.29). These genes were also enriched for GO Biological processes such as protein targeting to peroxisome (*ABCD2, PEX6, HACL1*) and positive regulation of neurogenesis (*CUX2, MME, RHOA, PLXNA4*) (Fig 3.11D, Fig 3.12B, Supp Table 3.25). Moreover, several genes in this set were associated with presenile dementia (*DNMT1, SPTLC1, SUCLA2, PRKAR1B, ALDH2, MME, VPS13A, IGF1, PLXNA4*) (Fig 3.11D, Fig 3.13B, Supp Table 3.27). In contrast, genes upregulated with age in Inh L3-5 SST ADGRG6 (88) were enriched for GO Biological processes such as positive regulation of signal transduction (*IRS1, GKAP1, SPIN1, TSPAN5, TM2D3, RBPJ*), and positive regulation of Notch signaling (*TSPAN5, TM2D3, RBPJ*) (Fig 3.11D, Fig 3.12A, Supp Table 3.24). In addition, they were associated with intelligence (*CDH2, PPA2, REEP3, GLCCI1, TSNARE1*) from the DisGeNET database (Fig 3.11D, Fig 3.13A, Supp Table 3.26).

Importantly, many of the enrichment terms did not reach significance after adjusting for multiple testing. Additionally, there were many genes which were expressed in a low proportion of nuclei (<20% of nuclei in the cell type under investigation) (Supp Table 3.23).

A


Enrichment analysis for genes upregulated with age: GO Biological Process
protein localization to plasma membrane (GO:0072659)
piecemeal microautophagy of the nucleus (GO:0034727)
lysosomal microautophagy (GO:0016237)
negative regulation of ERK1 and ERK2 cascade (GO:0070373)
regulation of TOR signaling (GO:0032006)
response to iron ion (GO:0010039)
protein localization to cell periphery (GO:1990778)
protein localization to membrane (GO:0072657)
peptidyl−threonine modification (GO:0018210)
negative regulation of immune effector process (GO:0002698)
T cell mediated immunity (GO:0002456)
autophagy of nucleus (GO:0044804)
supramolecular fiber organization (GO:0097435)
regulation of response to biotic stimulus (GO:0002831)
neuron projection morphogenesis (GO:0048812)
heterophilic cell−cell adhesion via plasma membrane cell adhesion molecules (GO:0007157)
fructose 2,6−bisphosphate metabolic process (GO:0006003)
diaphragm development (GO:0060539)
regulation of membrane tubulation (GO:1903525)
regulation of extracellular exosome assembly (GO:1903551)
peptide transport (GO:0015833)
autophagy of mitochondrion (GO:0000422)
aromatic amino acid transport (GO:0015801)
receptor−mediated endocytosis of virus by host cell (GO:0019065)
regulation of organelle assembly (GO:1902115)
negative regulation of TOR signaling (GO:0032007)
regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum (GO:0010880)
regulation of relaxation of cardiac muscle (GO:1901897)
regulation of alternative mRNA splicing, via spliceosome (GO:0000381)
cAMP−mediated signaling (GO:0019933)
receptor signaling pathway via STAT (GO:0097696)
positive regulation of signal transduction (GO:0009967)
positive regulation of Notch signaling pathway (GO:0045747)
reticulophagy (GO:0061709)
positive regulation of insulin receptor signaling pathway (GO:0046628)
positive regulation of cellular response to insulin stimulus (GO:1900078)
axonal transport of mitochondrion (GO:0019896)
regulation of platelet−derived growth factor receptor signaling pathway (GO:0010640)
regulation of lipase activity (GO:0060191)
membrane lipid biosynthetic process (GO:0046467)
padj=0.011; Number of genes=8
padj=0.014; Number of genes=2
padj=0.014; Number of genes=2
padj=0.021; Number of genes=5
padj=0.024; Number of genes=5
padj=0.033; Number of genes=3
padj=0.034; Number of genes=7
padj=0.034; Number of genes=8
padj=0.034; Number of genes=5
padj=0.045; Number of genes=2
padj=0.045; Number of genes=2
padj=0.053; Number of genes=2
padj=0.063; Number of genes=5
padj=0.075; Number of genes=2
padj=0.075; Number of genes=3
padj=0.075; Number of genes=2
padj=0.126; Number of genes=1
padj=0.126; Number of genes=1
padj=0.126; Number of genes=1
padj=0.126; Number of genes=1
padj=0.126; Number of genes=1
padj=0.13; Number of genes=2
padj=0.131; Number of genes=2
padj=0.131; Number of genes=2
padj=0.131; Number of genes=4
padj=0.158; Number of genes=3
padj=0.186; Number of genes=3
padj=0.186; Number of genes=2
padj=0.186; Number of genes=4
padj=0.186; Number of genes=3
padj=0.186; Number of genes=3
padj=0.197; Number of genes=6
padj=0.197; Number of genes=3
padj=0.197; Number of genes=2
padj=0.197; Number of genes=2
padj=0.197; Number of genes=2
padj=0.313; Number of genes=2
padj=0.313; Number of genes=2
padj=0.313; Number of genes=2
padj=0.313; Number of genes=3
Term
0.0
0.5
1.0
1.5
2.0
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

B

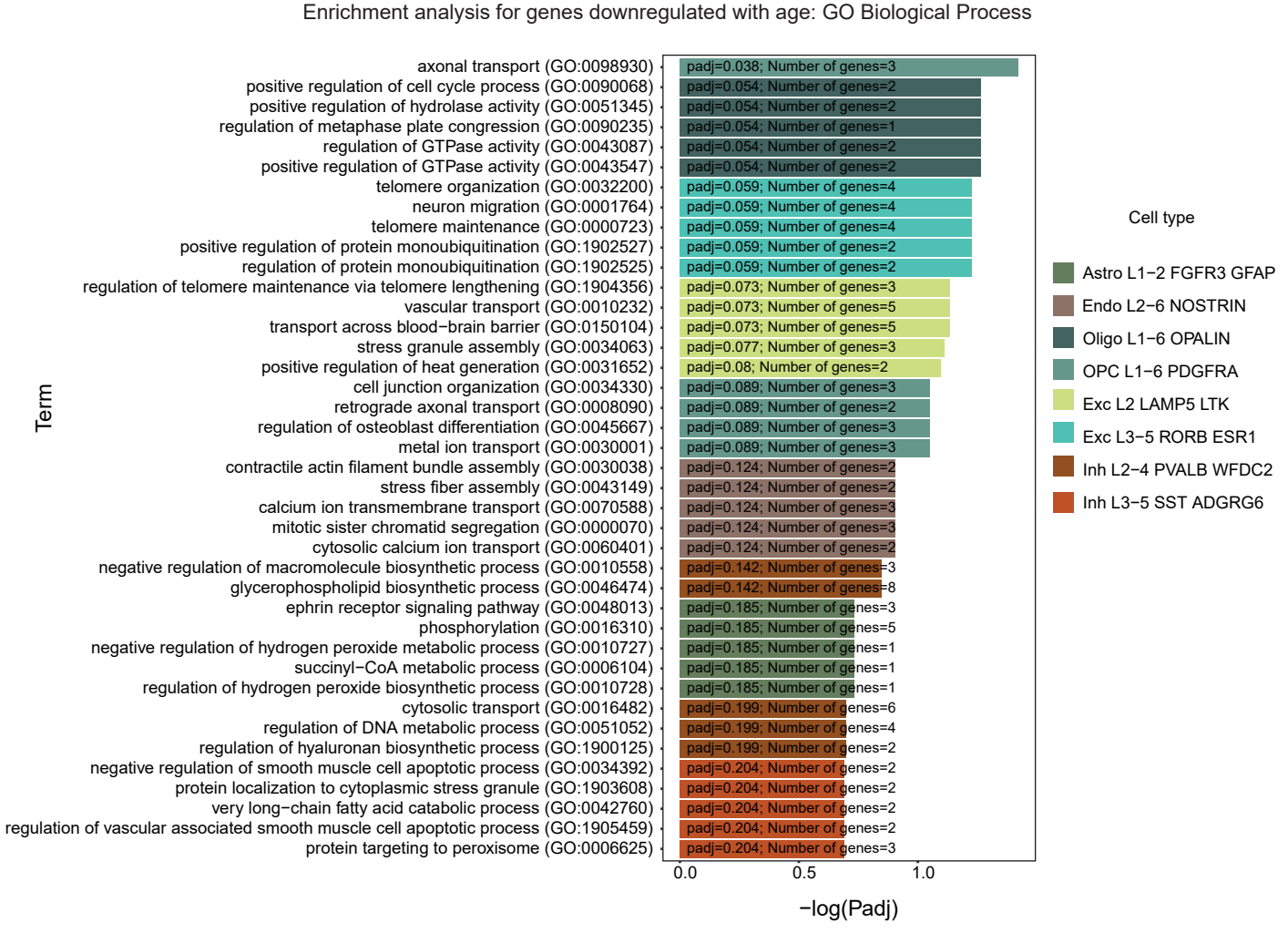
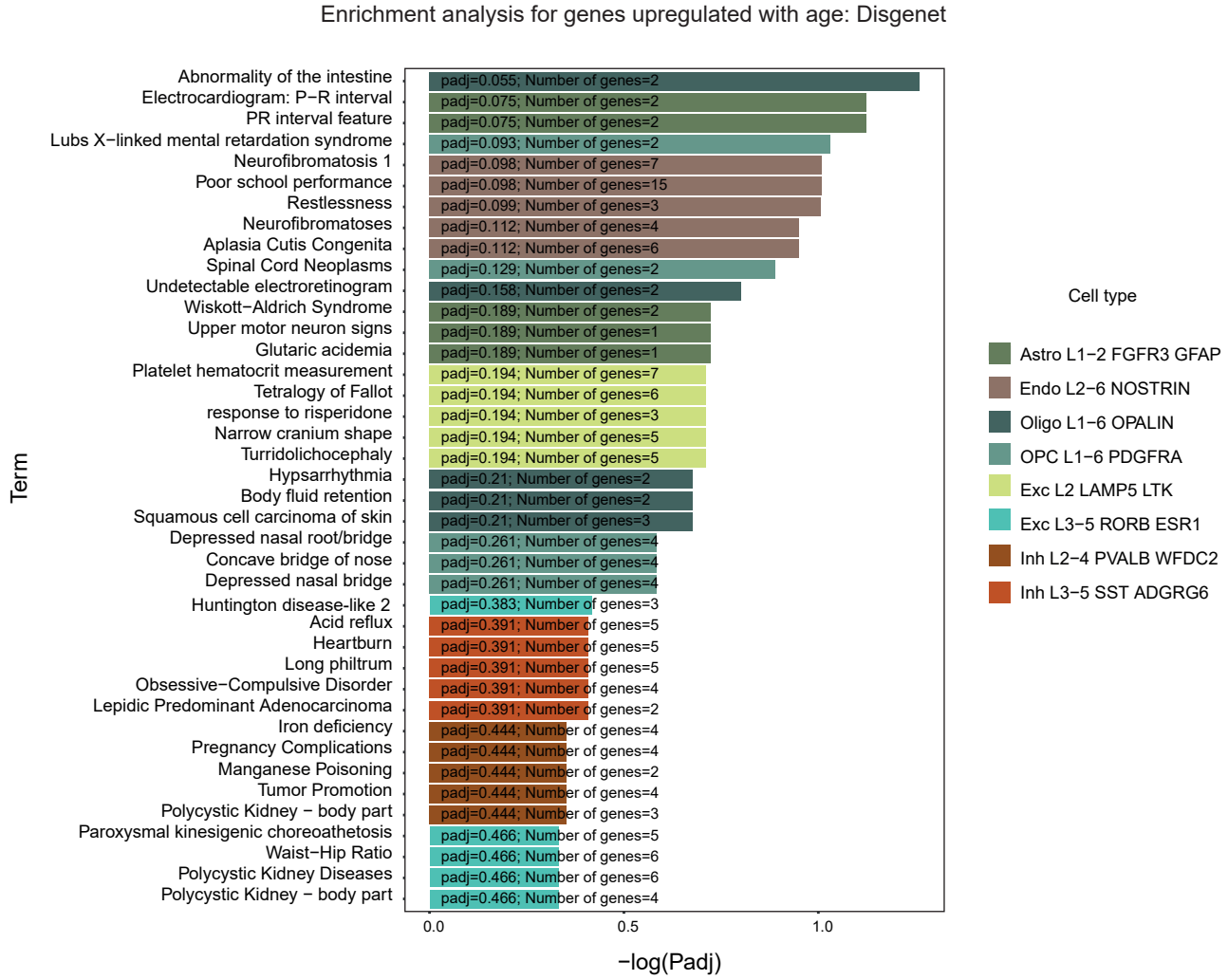

Enrichment analysis for genes downregulated with age: GO Biological Process
axonal transport (GO:0098930)
positive regulation of cell cycle process (GO:0090068)
positive regulation of hydrolase activity (GO:0051345)
regulation of metaphase plate congression (GO:0090235)
regulation of GTPase activity (GO:0043087)
positive regulation of GTPase activity (GO:0043547)
telomere organization (GO:0032200)
neuron migration (GO:0001764)
telomere maintenance (GO:0000723)
positive regulation of protein monoubiquitination (GO:1902527)
regulation of protein monoubiquitination (GO:1902525)
regulation of telomere maintenance via telomere lengthening (GO:1904356)
vascular transport (GO:0010232)
transport across blood−brain barrier (GO:0150104)
stress granule assembly (GO:0034063)
positive regulation of heat generation (GO:0031652)
cell junction organization (GO:0034330)
retrograde axonal transport (GO:0008090)
regulation of osteoblast differentiation (GO:0045667)
metal ion transport (GO:0030001)
contractile actin filament bundle assembly (GO:0030038)
stress fiber assembly (GO:0043149)
calcium ion transmembrane transport (GO:0070588)
mitotic sister chromatid segregation (GO:0000070)
cytosolic calcium ion transport (GO:0060401)
negative regulation of macromolecule biosynthetic process (GO:0010558)
glycerophospholipid biosynthetic process (GO:0046474)
ephrin receptor signaling pathway (GO:0048013)
phosphorylation (GO:0016310)
negative regulation of hydrogen peroxide metabolic process (GO:0010727)
succinyl−CoA metabolic process (GO:0006104)
regulation of hydrogen peroxide biosynthetic process (GO:0010728)
cytosolic transport (GO:0016482)
regulation of DNA metabolic process (GO:0051052)
regulation of hyaluronan biosynthetic process (GO:1900125)
negative regulation of smooth muscle cell apoptotic process (GO:0034392)
protein localization to cytoplasmic stress granule (GO:1903608)
very long−chain fatty acid catabolic process (GO:0042760)
regulation of vascular associated smooth muscle cell apoptotic process (GO:1905459)
protein targeting to peroxisome (GO:0006625)
padj=0.038; Number of genes=3
padj=0.054; Number of genes=2
padj=0.054; Number of genes=2
padj=0.054; Number of genes=1
padj=0.054; Number of genes=2
padj=0.054; Number of genes=2
padj=0.059; Number of genes=4
padj=0.059; Number of genes=4
padj=0.059; Number of genes=4
padj=0.059; Number of genes=2
padj=0.059; Number of genes=2
padj=0.073; Number of genes=3
padj=0.073; Number of genes=5
padj=0.073; Number of genes=5
padj=0.077; Number of genes=3
padj=0.08; Number of genes=2
padj=0.089; Number of genes=3
padj=0.089; Number of genes=2
padj=0.089; Number of genes=3
padj=0.089; Number of genes=3
padj=0.124; Number of genes=2
padj=0.124; Number of genes=2
padj=0.124; Number of genes=3
padj=0.124; Number of genes=3
padj=0.124; Number of genes=2
padj=0.142; Number of genes=3
padj=0.142; Number of genes=8
padj=0.185; Number of genes=3
padj=0.185; Number of genes=5
padj=0.185; Number of genes=1
padj=0.185; Number of genes=1
padj=0.185; Number of genes=1
padj=0.199; Number of genes=6
padj=0.199; Number of genes=4
padj=0.199; Number of genes=2
padj=0.204; Number of genes=2
padj=0.204; Number of genes=2
padj=0.204; Number of genes=2
padj=0.204; Number of genes=3
Term
0.0
0.5
1.0
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

Figure 3.12. Sets of up- and downregulated genes from the IDEAS analysis are implicated in various GO Biological processes. Putative biological processes that are upregulated (A) and downregulated (B) with age for Astro L1-2 FGFR3 GFAP, Endo L2-6 NOSTRIN, Oligo L1-6 OPALIN, OPC L1-6 PDGFRA, Exc L2 LAMP5 LTK, Exc L3-5 RORB ESR1, Inh L2-4 PVALB WFDC2, and Inh L3-5 SST ADGRG6 (coloured according to the MTG taxonomy[109]). The top 5 terms per cell type are displayed and ranked according to their p-values. The full enrichment results are in Supp Table 3.24 and Supp Table 3.25.

A


Enrichment analysis for genes upregulated with age: Disgenet
Abnormality of the intestine
Electrocardiogram: P−R interval
PR interval feature
Lubs X−linked mental retardation syndrome
Neurofibromatosis 1
Poor school performance
Restlessness
Neurofibromatoses
Aplasia Cutis Congenita
Spinal Cord Neoplasms
Undetectable electroretinogram
Wiskott−Aldrich Syndrome
Upper motor neuron signs
Glutaric acidemia
Platelet hematocrit measurement
Tetralogy of Fallot
response to risperidone
Narrow cranium shape
Turridolichocephaly
Hypsarrhythmia
Body fluid retention
Squamous cell carcinoma of skin
Depressed nasal root/bridge
Concave bridge of nose
Depressed nasal bridge
Huntington disease-like 2
Acid reflux
Heartburn
Long philtrum
Obsessive−Compulsive Disorder
Lepidic Predominant Adenocarcinoma
Iron deficiency
Pregnancy Complications
Manganese Poisoning
Tumor Promotion
Polycystic Kidney − body part
Paroxysmal kinesigenic choreoathetosis
Waist−Hip Ratio
Polycystic Kidney Diseases
Polycystic Kidney − body part
padj=0.055; Number of genes=2
padj=0.075; Number of genes=2
padj=0.075; Number of genes=2
padj=0.093; Number of genes=2
padj=0.098; Number of genes=7
padj=0.098; Number of genes=15
padj=0.099; Number of genes=3
padj=0.112; Number of genes=4
padj=0.112; Number of genes=6
padj=0.129; Number of genes=2
padj=0.158; Number of genes=2
padj=0.189; Number of genes=2
padj=0.189; Number of genes=1
padj=0.189; Number of genes=1
padj=0.194; Number of genes=7
padj=0.194; Number of genes=6
padj=0.194; Number of genes=3
padj=0.194; Number of genes=5
padj=0.194; Number of genes=5
padj=0.21; Number of genes=2
padj=0.21; Number of genes=2
padj=0.21; Number of genes=3
padj=0.261; Number of genes=4
padj=0.261; Number of genes=4
padj=0.261; Number of genes=4
padj=0.383; Number of genes=3
padj=0.391; Number of genes=5
padj=0.391; Number of genes=5
padj=0.391; Number of genes=5
padj=0.391; Number of genes=4
padj=0.391; Number of genes=2
padj=0.444; Number of genes=4
padj=0.444; Number of genes=4
padj=0.444; Number of genes=2
padj=0.444; Number of genes=4
padj=0.444; Number of genes=3
padj=0.466; Number of genes=5
padj=0.466; Number of genes=6
padj=0.466; Number of genes=6
padj=0.466; Number of genes=4
Term
0.0
0.5
1.0
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

B

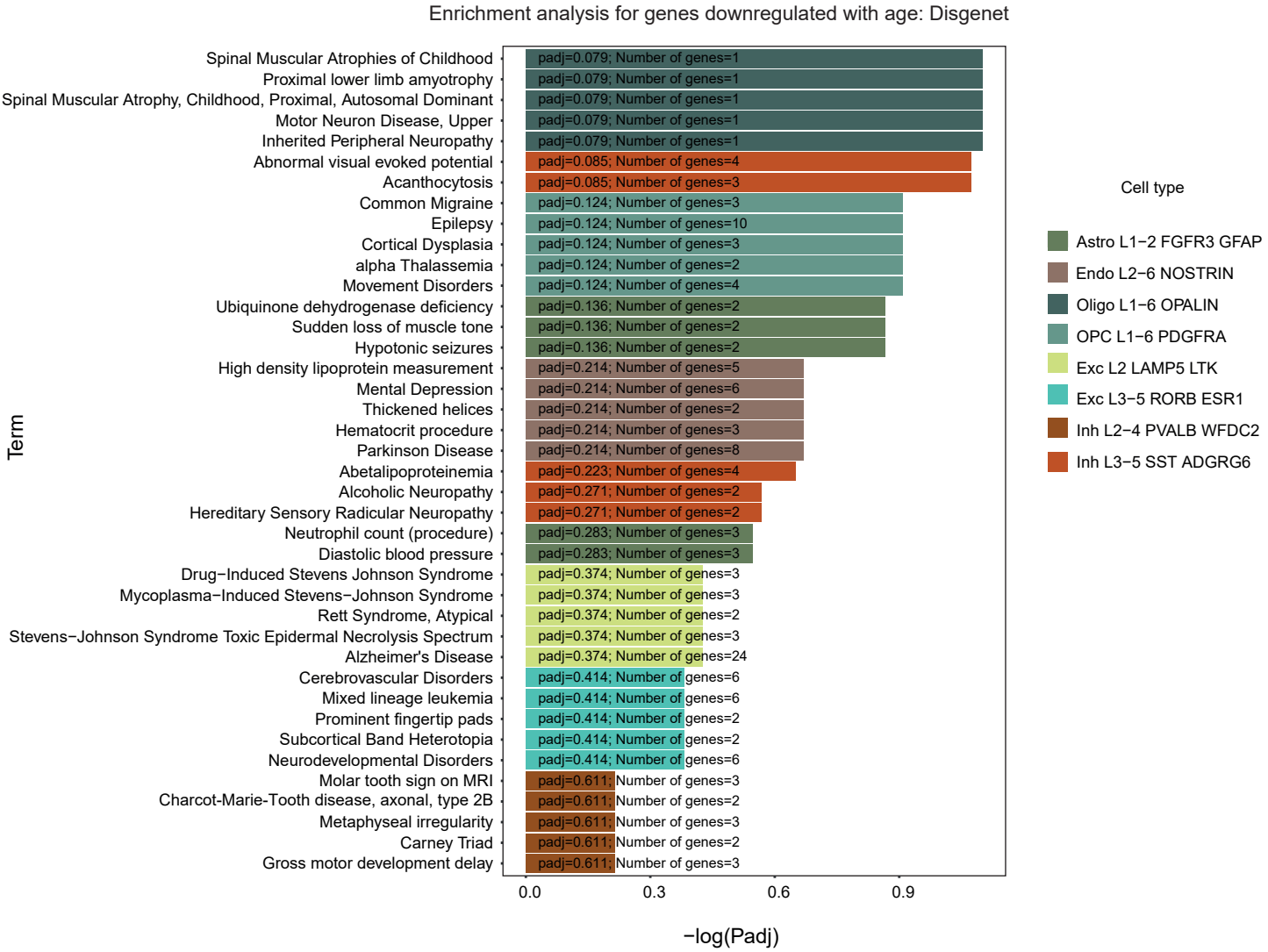
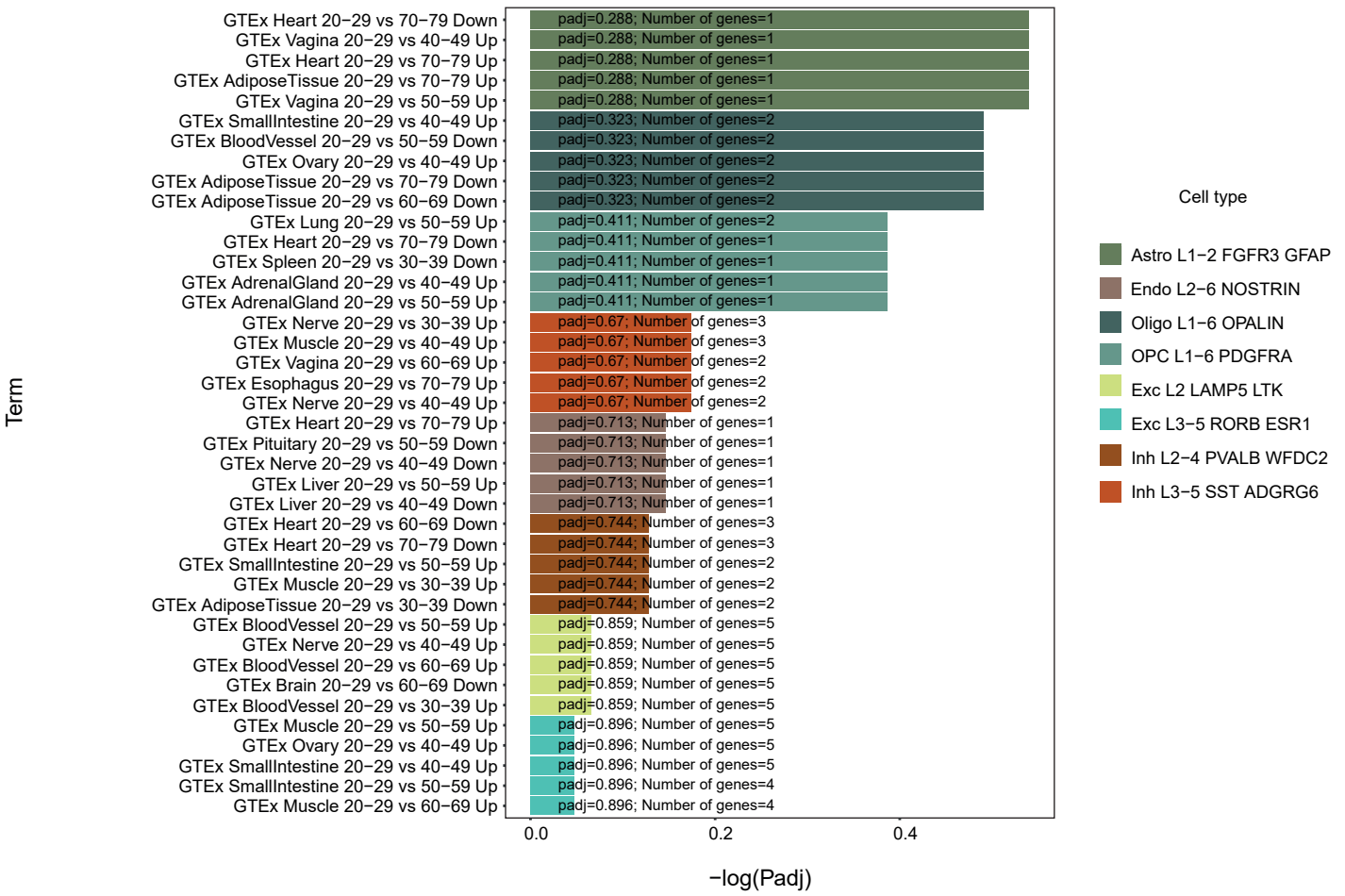

Enrichment analysis for genes downregulated with age: Disgenet
Spinal Muscular Atrophies of Childhood
Proximal lower limb amyotrophy
Spinal Muscular Atrophy, Childhood, Proximal, Autosomal Dominant
Motor Neuron Disease, Upper
Inherited Peripheral Neuropathy
Abnormal visual evoked potential
Acanthocytosis
Common Migraine
Epilepsy
Cortical Dysplasia
alpha Thalassemia
Movement Disorders
Ubiquinone dehydrogenase deficiency
Sudden loss of muscle tone
Hypotonic seizures
High density lipoprotein measurement
Mental Depression
Thickened helices
Hematocrit procedure
Parkinson Disease
Abetalipoproteinemia
Alcoholic Neuropathy
Hereditary Sensory Radicular Neuropathy
Neutrophil count (procedure)
Diastolic blood pressure
Drug−Induced Stevens Johnson Syndrome
Mycoplasma−Induced Stevens−Johnson Syndrome
Rett Syndrome, Atypical
Stevens−Johnson Syndrome Toxic Epidermal Necrolysis Spectrum
Alzheimer's Disease
Cerebrovascular Disorders
Mixed lineage leukemia
Prominent fingertip pads
Subcortical Band Heterotopia
Neurodevelopmental Disorders
Molar tooth sign on MRI
Charcot-Marie-Tooth disease, axonal, type 2B
Metaphyseal irregularity
Carney Triad
Gross motor development delay
padj=0.079; Number of genes=1
padj=0.079; Number of genes=1
padj=0.079; Number of genes=1
padj=0.079; Number of genes=1
padj=0.079; Number of genes=1
padj=0.085; Number of genes=4
padj=0.085; Number of genes=3
padj=0.124; Number of genes=3
padj=0.124; Number of genes=10
padj=0.124; Number of genes=3
padj=0.124; Number of genes=2
padj=0.124; Number of genes=4
padj=0.136; Number of genes=2
padj=0.136; Number of genes=2
padj=0.136; Number of genes=2
padj=0.214; Number of genes=5
padj=0.214; Number of genes=6
padj=0.214; Number of genes=2
padj=0.214; Number of genes=3
padj=0.214; Number of genes=8
padj=0.223; Number of genes=4
padj=0.271; Number of genes=2
padj=0.271; Number of genes=2
padj=0.283; Number of genes=3
padj=0.283; Number of genes=3
padj=0.374; Number of genes=3
padj=0.374; Number of genes=3
padj=0.374; Number of genes=2
padj=0.374; Number of genes=3
padj=0.374; Number of genes=24
padj=0.414; Number of genes=6
padj=0.414; Number of genes=6
padj=0.414; Number of genes=2
padj=0.414; Number of genes=2
padj=0.414; Number of genes=6
padj=0.611; Number of genes=3
padj=0.611; Number of genes=2
padj=0.611; Number of genes=3
padj=0.611; Number of genes=2
padj=0.611; Number of genes=3
Term
0.0
0.3
0.6
0.9
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

Figure 3.13. Sets of up and downregulated genes from the IDEAS analysis are implicated in various diseases from the DisGeNET database. Putative diseases or disease-related processes that are upregulated (A) and downregulated (B) with age in Astro L1-2 FGFR3 GFAP, Endo L2-6 NOSTRIN, Oligo L1-6 OPALIN, OPC L1-6 PDGFRA, Exc L2 LAMP5 LTK, Exc L3-5 RORB ESR1, Inh L2-4 PVALB WFDC2, and Inh L3-5 SST ADGRG6 (coloured according to the MTG taxonomy[109]). The top 5 terms per cell type are displayed and ranked according to their p-values. The full enrichment results are in Supp Table 3.26 and Supp Table 3.27.

A


Enrichment analysis for genes upregulated with age: GTEx Aging Signatures
Term
GTEx Heart 20−29 vs 70−79 Down
GTEx Vagina 20−29 vs 40−49 Up
GTEx Heart 20−29 vs 70−79 Up
GTEx AdiposeTissue 20−29 vs 70−79 Up
GTEx Vagina 20−29 vs 50−59 Up
GTEx SmallIntestine 20−29 vs 40−49 Up
GTEx BloodVessel 20−29 vs 50−59 Down
GTEx Ovary 20−29 vs 40−49 Up
GTEx AdiposeTissue 20−29 vs 70−79 Down
GTEx AdiposeTissue 20−29 vs 60−69 Down
GTEx Lung 20−29 vs 50−59 Up
GTEx Heart 20−29 vs 70−79 Down
GTEx Spleen 20−29 vs 30−39 Down
GTEx AdrenalGland 20−29 vs 40−49 Up
GTEx AdrenalGland 20−29 vs 50−59 Up
GTEx Nerve 20−29 vs 30−39 Up
GTEx Muscle 20−29 vs 40−49 Up
GTEx Vagina 20−29 vs 60−69 Up
GTEx Esophagus 20−29 vs 70−79 Up
GTEx Nerve 20−29 vs 40−49 Up
GTEx Heart 20−29 vs 70−79 Up
GTEx Pituitary 20−29 vs 50−59 Down
GTEx Nerve 20−29 vs 40−49 Down
GTEx Liver 20−29 vs 50−59 Up
GTEx Liver 20−29 vs 40−49 Down
GTEx Heart 20−29 vs 60−69 Down
GTEx Heart 20−29 vs 70−79 Down
GTEx SmallIntestine 20−29 vs 50−59 Up
GTEx Muscle 20−29 vs 30−39 Up
GTEx AdiposeTissue 20−29 vs 30−39 Down
GTEx BloodVessel 20−29 vs 50−59 Up
GTEx Nerve 20−29 vs 40−49 Up
GTEx BloodVessel 20−29 vs 60−69 Up
GTEx Brain 20−29 vs 60−69 Down
GTEx BloodVessel 20−29 vs 30−39 Up
GTEx Muscle 20−29 vs 50−59 Up
GTEx Ovary 20−29 vs 40−49 Up
GTEx SmallIntestine 20−29 vs 40−49 Up
GTEx SmallIntestine 20−29 vs 50−59 Up
GTEx Muscle 20−29 vs 60−69 Up
padj=0.288; Number of genes=1
padj=0.323; Number of genes=2
padj=0.411; Number of genes=2
padj=0.411; Number of genes=1
padj=0.67; Number of genes=3
padj=0.67; Number of genes=2
padj=0.713; Number of genes=1
padj=0.744; Number of genes=3
padj=0.744; Number of genes=2
padj=0.859; Number of genes=5
padj=0.896; Number of genes=5
padj=0.896; Number of genes=4
0.0
0.2
0.4
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

B

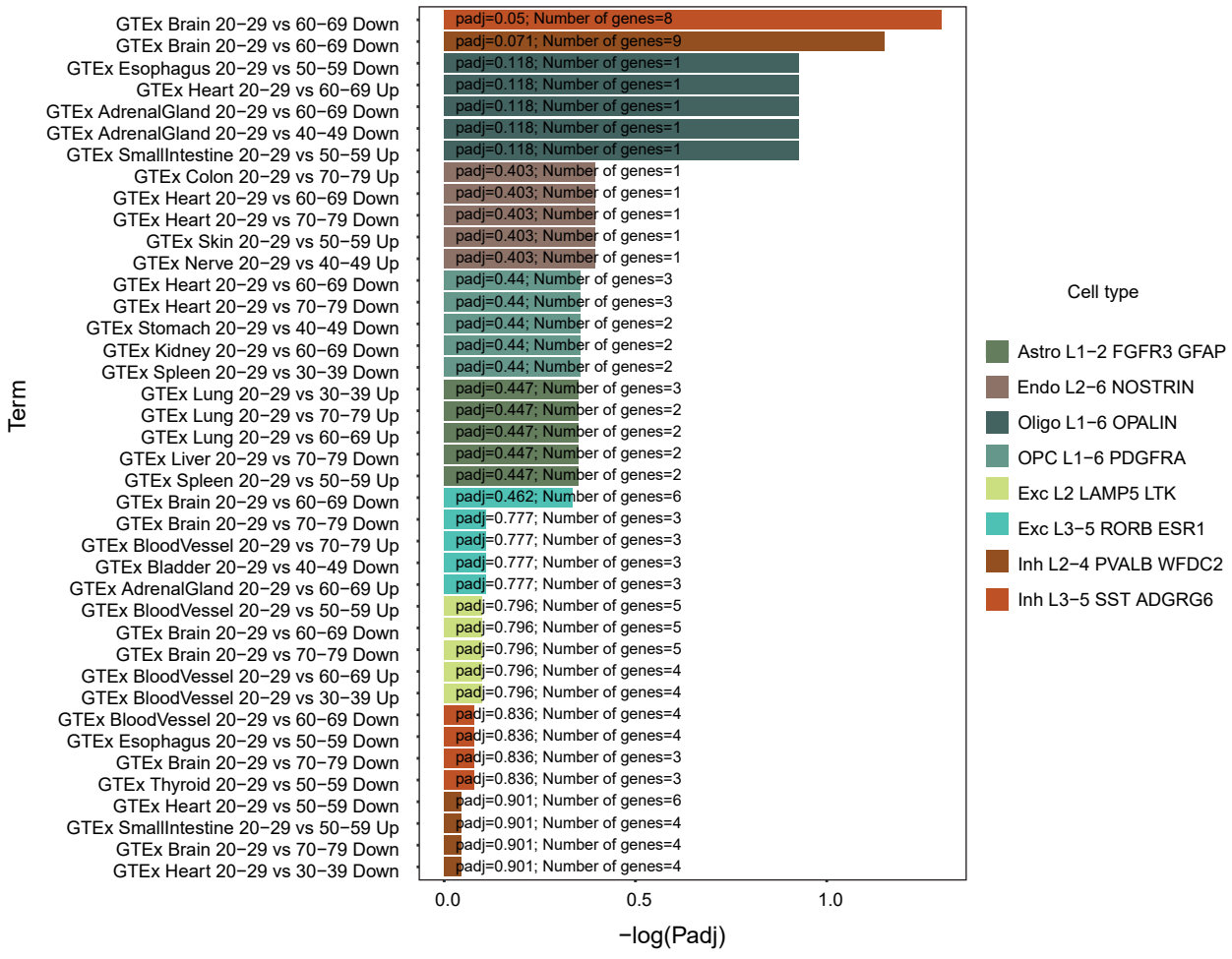

Enrichment analysis for genes downregulated with age: GTEx Aging Signatures
Term
GTEx Brain 20−29 vs 60−69 Down
GTEx Brain 20−29 vs 60−69 Down
GTEx Esophagus 20−29 vs 50−59 Down
GTEx Heart 20−29 vs 60−69 Up
GTEx AdrenalGland 20−29 vs 60−69 Down
GTEx AdrenalGland 20−29 vs 40−49 Down
GTEx SmallIntestine 20−29 vs 50−59 Up
GTEx Colon 20−29 vs 70−79 Up
GTEx Heart 20−29 vs 60−69 Down
GTEx Heart 20−29 vs 70−79 Down
GTEx Skin 20−29 vs 50−59 Up
GTEx Nerve 20−29 vs 40−49 Up
GTEx Heart 20−29 vs 60−69 Down
GTEx Heart 20−29 vs 70−79 Down
GTEx Stomach 20−29 vs 40−49 Down
GTEx Kidney 20−29 vs 60−69 Down
GTEx Spleen 20−29 vs 30−39 Down
GTEx Lung 20−29 vs 30−39 Up
GTEx Lung 20−29 vs 70−79 Up
GTEx Lung 20−29 vs 60−69 Up
GTEx Liver 20−29 vs 70−79 Down
GTEx Spleen 20−29 vs 50−59 Up
GTEx Brain 20−29 vs 60−69 Down
GTEx Brain 20−29 vs 70−79 Down
GTEx BloodVessel 20−29 vs 70−79 Up
GTEx Bladder 20−29 vs 40−49 Down
GTEx AdrenalGland 20−29 vs 60−69 Up
GTEx BloodVessel 20−29 vs 50−59 Up
GTEx Brain 20−29 vs 60−69 Down
GTEx Brain 20−29 vs 70−79 Down
GTEx BloodVessel 20−29 vs 60−69 Up
GTEx BloodVessel 20−29 vs 30−39 Up
GTEx BloodVessel 20−29 vs 60−69 Down
GTEx Esophagus 20−29 vs 50−59 Down
GTEx Brain 20−29 vs 70−79 Down
GTEx Thyroid 20−29 vs 50−59 Down
GTEx Heart 20−29 vs 50−59 Down
GTEx SmallIntestine 20−29 vs 50−59 Up
GTEx Brain 20−29 vs 70−79 Down
GTEx Heart 20−29 vs 30−39 Down
padj=0.05; Number of genes=8
padj=0.071; Number of genes=9
padj=0.118; Number of genes=1
padj=0.403; Number of genes=1
padj=0.44; Number of genes=3
padj=0.44; Number of genes=2
padj=0.447; Number of genes=3
padj=0.447; Number of genes=2
padj=0.462; Number of genes=6
padj=0.777; Number of genes=3
padj=0.796; Number of genes=5
padj=0.796; Number of genes=4
padj=0.836; Number of genes=4
padj=0.836; Number of genes=3
padj=0.901; Number of genes=6
padj=0.901; Number of genes=4
0.0
0.5
1.0
−log(Padj)
Cell type
Astro L1−2 FGFR3 GFAP
Endo L2−6 NOSTRIN
Oligo L1−6 OPALIN
OPC L1−6 PDGFRA
Exc L2 LAMP5 LTK
Exc L3−5 RORB ESR1
Inh L2−4 PVALB WFDC2
Inh L3−5 SST ADGRG6

Figure 3.14. Sets of up- and downregulated genes from the IDEAS analysis are associated with the GTEx Aging Signatures 2021 database. Upregulated (A) and downregulated (B) DEGs associated with terms from the GTEx Aging Signatures 2021 database for Astro L1-2 FGFR3 GFAP, Endo L2-6 NOSTRIN, Oligo L1-6 OPALIN, OPC L1-6 PDGFRA, Exc L2 LAMP5 LTK, Exc L3-5 RORB ESR1, Inh L2-4 PVALB WFDC2, and Inh L3-5 SST ADGRG6 (coloured according to the MTG taxonomy[109]). The top 5 terms per cell type are displayed and ranked according to their p-values. The full enrichment results are in Supp Table 3.28 and Supp Table 3.29.

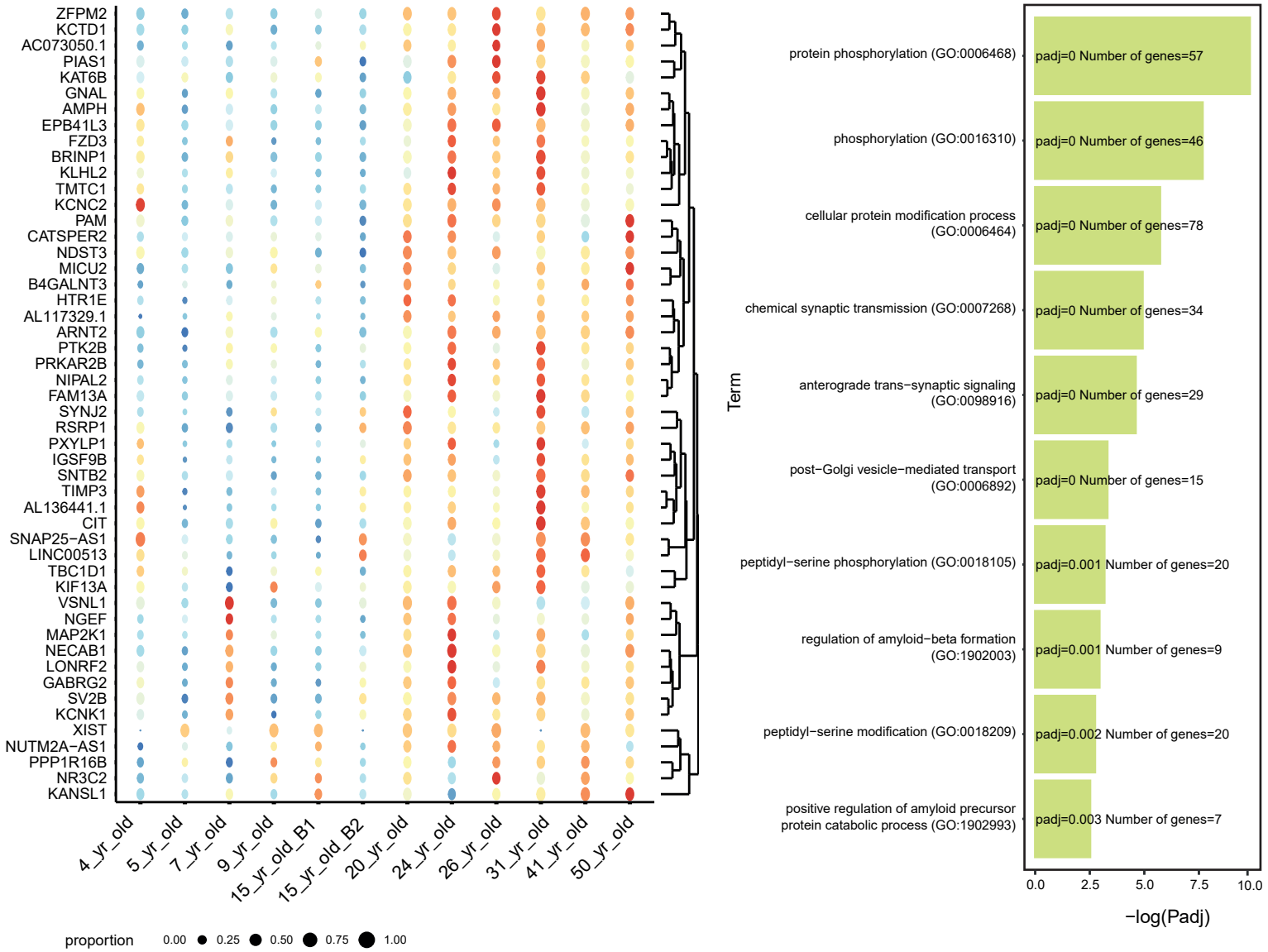
3.8. Changes in the proportion of nuclei expressing each gene with age

Unlike bulk RNA-seq data, snRNA-seq data presents the opportunity to examine the proportion of nuclei expressing each gene within a given cell type. This could be useful for identifying genes that may not change in expression level within a given cell type over time but show a change in the number of cells expressing them within that cell type as the brain matures. Alternatively, if the sequencing saturation is low, a change in the proportion of nuclei expressing a gene may be indicative of a change in the level of expression since genes which are more highly expressed should in theory have a greater likelihood of being sampled. Thus, assessing the change in the proportion of nuclei expressing a gene with age may provide insight into genes which are relevant to the process of brain maturation.

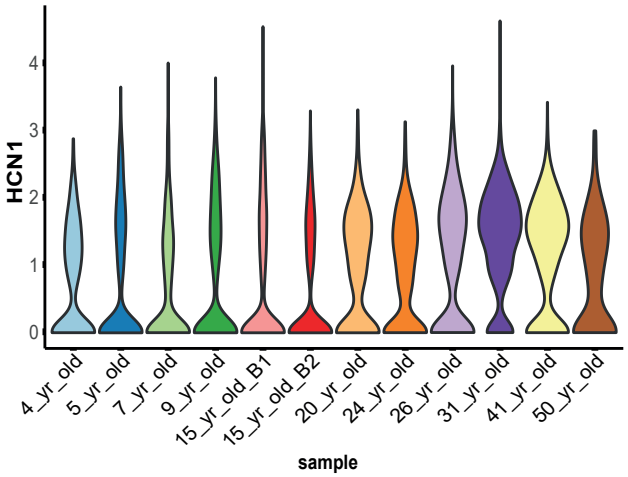
To this end, a hypothesis test was performed to compare the proportion of nuclei expressing each gene per cell type between the adult and paediatric samples. This revealed a significant difference for 5440 genes in the Exc L2 LAMP5 LTK and 1 gene in Endo L2-6 NOSTRIN, whereas no significant difference was observed for any other cell types (Supp Table 3.30). Notably, it appeared that for all the significant genes identified in Exc L2 LAMP5 LTK there was an increase in the proportion of nuclei expressing each gene in the adult datasets compared to the paediatric datasets. This increasing trend in adult samples versus paediatric samples remained after using an alternative method to account for the total number of nuclei per group, by down-sampling to equivalent numbers of nuclei (Supp Table 3.30). Using the down sampling approach, a total of 5192 genes showed a significant change in the number of Exc L2 LAMP5 LTK neurons expressing them, and 4680 genes were shared with the original proportion analysis method (Supp Table 3.30). The down-sampling approach also identified a significant difference in the number of nuclei expressing one gene, *SMS,* in Exc L3-5 RORB ESR1 (Supp Table 3.30). Based on the proportion analysis method, the genes that were expressed by at least 20% more Exc L2 LAMP5 LTK nuclei in the adult datasets versus the paediatric datasets were highly enriched for several GO biological processes such as protein phosphorylation, synaptic signalling, vesicle-mediated transport, and amyloid beta formation (Fig 3.15A, Supp Table 3.31). The significant genes included *HCN1* (Fig 3.15B) and *SCN1A* (Fig 3.15C) which are channels involved in regulating excitability in neurons[274,275]. These were not found to be differentially expressed by IDEAS in Exc L2 LAMP5 LTK (Fig 3.11A, Supp Table 3.23) despite showing a clear increase in both the proportion of nuclei expressing the gene and relative level of expression between paediatrics and adults (Fig 3.15A-C).

A

Exc L2 LAMP5 LTK



B



C

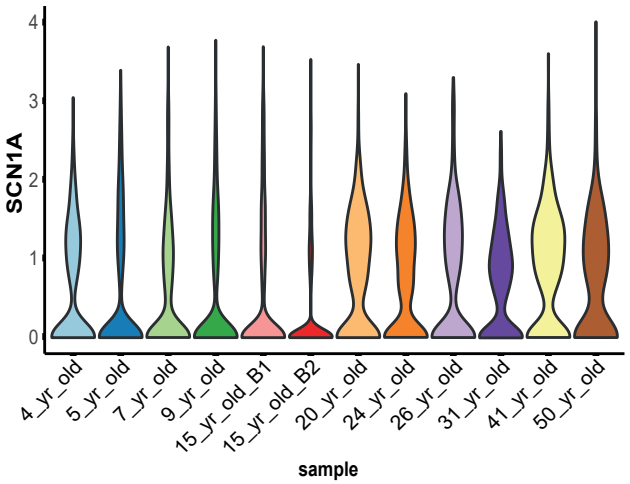


Figure 3.15. Genes expressed by a significantly different proportion of paediatric and adult Exc L2 LAMP5 LTK nuclei. (A) Top 50 significant genes ranked according to the difference in the percentage of nuclei expressing them between adult and paediatric samples (see Supp Table 3.30). Data points are coloured according to the level of expression (scaled average normalised gene counts). The size of the dots indicates the proportion of nuclei expressing the gene per sample (Left). Enrichment plots showing the top 10 enriched terms (y-axis) ranked by p-value (x-axis) which are associated with the genes expressed by at least 20% more nuclei in the adult datasets versus the paediatric datasets. The GO Biological Processes 2021 database from the Enrichr package was used to perform GSEA. Plots are coloured according to the MTG taxonomy from Hodge et al. (2019) (Right). (B, C) Violin plots showing the level and proportion of nuclei expressing *HCN1* (B) and *SCN1A* (C) across the 12 samples in Exc L2 LAMP5 LTK. Counts are the normalised gene expression counts.

3.9. Analysis of long non-coding RNAs for two genes of interest

Over the past decade, the regulatory functions of numerous lncRNAs in the brain have been begun to be described including roles in brain cell proliferation[155,156], neuronal differentiation[149,157,158], and synaptic plasticity[160]. Additionally, there is evidence supporting a role for lncRNAs in neurological disorders, including schizophrenia[161], autism[162], and epilepsy[163161–163], yet the functions of the majority of lncRNAs remain unknown. Here I use computational means to probe the possible cell type-specific functions of two lncRNAs of interest from the snRNA-seq analysis, namely *LINC00499* and *AC004852.2*. Both were identified as markers of a particular cell type by NS-Forest and were differentially expressed with age within a particular cell type. *LINC00499* was identified as a novel marker of Astro L1-6 FGFR3 SLC14A1 and was differentially expressed in this cell type in both the DESeq2 and Psupertime analysis. It resides on the forward strand of chromosome 4 and is 541 nucleotides in length. *AC004852.2*, which is 878 nucleotides long and found on chromosome 7, was identified as a novel marker of OPC L1-6 PDGFRA and was temporally regulated according to Psupertime.

One strategy to investigate putative functions of lncRNAs is to discover which genes their expression is correlated with and perform GSEA on the top co-expressed protein coding genes with known functions – a method described as 'Guilt by association analysis' (GBA)[259]. To this end I used a Bayesian correlation method[260] to obtain a gene-gene similarity matrix for each of the cell types under investigation using the gene by cell matrices from the snRNA-seq datasets as input. The co-expressed protein-coding genes corresponding to the lncRNA of interest were extracted and ranked according to their correlation score. GSEA was performed on two separate lists of the top 50 positively and top 50 negatively correlated genes (Supp Table 3.32). For *LINC00499,* the top 50 positively correlated genes were enriched for terms relating to immune functioning as well as terms such as substantia nigra development and positive regulation of gene expression (Fig 3.16A-B). On the other hand, the top 50 negatively correlated genes were enriched for terms relating to synapse assembly and function (Fig 3.16A-B). Notably, the top 50 positively and negatively co-expressed genes appeared to show a pattern according to batch, with the publicly available datasets being more similar to each other than to our datasets (Fig 3.16A-B). Additionally, many of the co-expressed genes were expressed in a low percentage of nuclei in Astro L1-6 FGFR3 SLC14A1 (Fig 3.16A). In contrast, *LINC00499* was expressed in 61% of nuclei in this cell type across the 12 samples (Fig 3.16A, Supp Table 3.32). Nevertheless, there were

several genes which were expressed in at least 35% of nuclei in Astro L1-6 FGFR3 SLC14A1 (Supp Table 3.32) and showed either a distinct agreeing or opposing expression pattern to that of *LINC00499,* including, *AQP4, PDE3B, FAM171B, TENM4, NLGN1, GRID2, RGS6, RGS7,* (Supp Fig 3.30A, Supp Table 3.32).

The top 50 positively co-expressed genes with *AC004852.2* were associated with terms such as synapse organization, anterograde trans-synaptic signaling, action potential, and cell-cell adhesion whereas the top 50 negatively co-expressed genes were enriched for terms such as myelination, axonogenesis, and positive regulation of phosphorylation (Fig 3.17A-B). These genes included *SCN1A, CHL1,* and *NOVA1* (Fig 3.17A, Supp Fig 3.30B). Markedly, the majority of genes in the list of top 50 negatively correlated genes were expressed in fewer than 20% of nuclei in OPC L1-6 PDGFRA, with exceptions being *FRMD4A, FMNL2*, and *MBP* (Fig 3.17A, Supp Table 3.32). In contrast, most of the top 50 positively correlated genes were expressed in at least 50% of nuclei, with half of them expressed in more than 90% of nuclei in OPC L1-6 PDGFRA (Fig 3.17A, Supp Table 3.32). This is similar to *AC004852.2* which was expressed in 73% of OPC L1-6 PDGFRA nuclei (Supp Table 3.32).

In addition to guilt by association analysis, determining the interaction partners of lncRNAs is an alternate strategy to explore the functions of these molecules without having to do functional genetic screens. To investigate putative DNA binding motifs for the selected lncRNAs, the sequence of each molecule was queried against regions near promoters across the entire genome using FasimTarget[262], which computationally tests all known base pairing rules required to form RNA:DNA triplexes. The list of DNA binding motif hits were ranked according to the percentage of nuclei expressing the corresponding gene and the top 50 hits used in GSEA (Supp Table 3.32). Many of the genes on the list of putative DNA interaction partners for *LINC00499* showed an opposing expression trajectory with age to that of *LINC00499,* which resembled the pattern observed by the top 50 negatively correlated genes from the GBA (Fig 3.16C). An intersection of the top 50 positively and top 50 negatively correlated genes from the GBA with the top 50 LongTarget genes identified *RGS7* as a shared gene between these analyses (Fig 3.16A, Fig 3.16C, Supp Fig 3.30A, Supp Table 3.32). The list of FasimTarget hits for *LINC00499* was enriched for terms such as positive regulation of Wnt signaling, forebrain neuron development, and apoptotic process (Fig 3.16D).

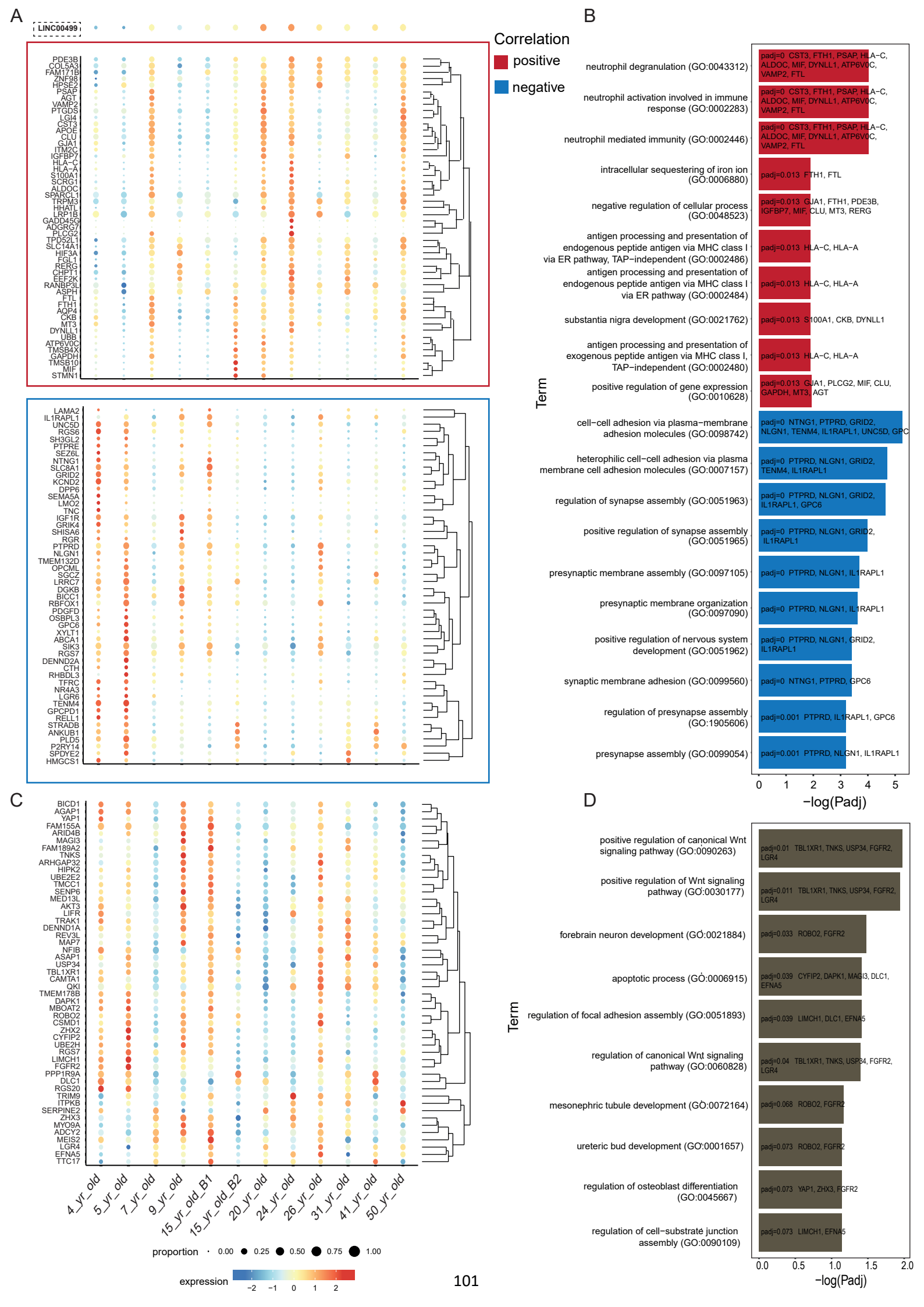

A
LINC00499
PDE3B
COL5A3
FAM171B
ZNF98
HPSE2
PSAP
AGT
VAMP2
PTGDS
LGI4
CST3
APOE
CLU
GJA1
ITM2C
IGFBP7
HLA−C
HLA−A
S100A1
SCRG1
ALDOC
SPARCL1
TRPM3
HHATL
LRP1B
GADD45G
ADGRG7
PLCG2
TPD52L1
SLC14A1
HIF3A
FGL1
RERG
CHPT1
EEF2K
RANBP3L
ASPH
FTL
FTH1
AQP4
CKB
MT3
DYNLL1
UBB
ATP6V0C
TMSB4X
GAPDH
TMSB10
MIF
STMN1
LAMA2
IL1RAPL1
UNC5D
RGS6
SH3GL2
PTPRE
SEZ6L
NTNG1
SLC8A1
GRID2
KCND2
DPP6
SEMA5A
LMO2
TNC
IGF1R
GRIK4
SHISA6
RGR
PTPRD
NLGN1
TMEM132D
OPCML
SGCZ
LRRC7
DGKB
BICC1
RBFOX1
PDGFD
OSBPL3
GPC6
XYLT1
ABCA1
SIK3
RGS7
DENND2A
CTH
RHBDL3
TFRC
NR4A3
LGR6
TENM4
GPCPD1
RELL1
STRADB
ANKUB1
PLD5
P2RY14
SPDYE2
HMGCS1
Correlation
positive
negative
B
neutrophil degranulation (GO:0043312)
padj=0 CST3, FTH1, PSAP, HLA−C, ALDOC, MIF, DYNLL1, ATP6V0C, VAMP2, FTL
neutrophil activation involved in immune response (GO:0002283)
padj=0 CST3, FTH1, PSAP, HLA−C, ALDOC, MIF, DYNLL1, ATP6V0C, VAMP2, FTL
neutrophil mediated immunity (GO:0002446)
padj=0 CST3, FTH1, PSAP, HLA−C, ALDOC, MIF, DYNLL1, ATP6V0C, VAMP2, FTL
intracellular sequestering of iron ion (GO:0006880)
padj=0.013 FTH1, FTL
negative regulation of cellular process (GO:0048523)
padj=0.013 GJA1, FTH1, PDE3B, IGFBP7, MIF, CLU, MT3, RERG
antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP−independent (GO:0002486)
padj=0.013 HLA−C, HLA−A
antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway (GO:0002484)
padj=0.013 HLA−C, HLA−A
substantia nigra development (GO:0021762)
padj=0.013 S100A1, CKB, DYNLL1
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP−independent (GO:0002480)
padj=0.013 HLA−C, HLA−A
positive regulation of gene expression (GO:0010628)
padj=0.013 GJA1, PLCG2, MIF, CLU, GAPDH, MT3, AGT
cell−cell adhesion via plasma−membrane adhesion molecules (GO:0098742)
padj=0 NTNG1, PTPRD, GRID2, NLGN1, TENM4, IL1RAPL1, UNC5D, GPC
heterophilic cell−cell adhesion via plasma membrane cell adhesion molecules (GO:0007157)
padj=0 PTPRD, NLGN1, GRID2, TENM4, IL1RAPL1
regulation of synapse assembly (GO:0051963)
padj=0 PTPRD, NLGN1, GRID2, IL1RAPL1, GPC6
positive regulation of synapse assembly (GO:0051965)
padj=0 PTPRD, NLGN1, GRID2, IL1RAPL1
presynaptic membrane assembly (GO:0097105)
padj=0 PTPRD, NLGN1, IL1RAPL1
presynaptic membrane organization (GO:0097090)
padj=0 PTPRD, NLGN1, IL1RAPL1
positive regulation of nervous system development (GO:0051962)
padj=0 PTPRD, NLGN1, GRID2, IL1RAPL1
synaptic membrane adhesion (GO:0099560)
padj=0 NTNG1, PTPRD, GPC6
regulation of presynapse assembly (GO:1905606)
padj=0.001 PTPRD, IL1RAPL1, GPC6
presynapse assembly (GO:0099054)
padj=0.001 PTPRD, NLGN1, IL1RAPL1
Term
0 1 2 3 4 5
−log(Padj)
C
BICD1
AGAP1
YAP1
FAM155A
ARID4B
MAGI3
FAM189A2
TNKS
ARHGAP32
HIPK2
UBE2E2
TMCC1
SENP6
MED13L
AKT3
LIFR
TRAK1
DENND1A
REV3L
MAP7
NFIB
ASAP1
USP34
TBL1XR1
CAMTA1
QKI
TMEM178B
DAPK1
MBOAT2
ROBO2
CSMD1
ZHX2
CYFIP2
UBE2H
RGS7
LIMCH1
FGFR2
PPP1R9A
DLC1
RGS20
TRIM9
ITPKB
SERPINE2
ZHX3
MYO9A
ADCY2
MEIS2
LGR4
EFNA5
TTC17
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
proportion 0.00 0.25 0.50 0.75 1.00
expression −2 −1 0 1 2
D
positive regulation of canonical Wnt signaling pathway (GO:0090263)
padj=0.01 TBL1XR1, TNKS, USP34, FGFR2, LGR4
positive regulation of Wnt signaling pathway (GO:0030177)
padj=0.011 TBL1XR1, TNKS, USP34, FGFR2, LGR4
forebrain neuron development (GO:0021884)
padj=0.033 ROBO2, FGFR2
apoptotic process (GO:0006915)
padj=0.039 CYFIP2, DAPK1, MAGI3, DLC1, EFNA5
regulation of focal adhesion assembly (GO:0051893)
padj=0.039 LIMCH1, DLC1, EFNA5
regulation of canonical Wnt signaling pathway (GO:0060828)
padj=0.04 TBL1XR1, TNKS, USP34, FGFR2, LGR4
mesonephric tubule development (GO:0072164)
padj=0.068 ROBO2, FGFR2
ureteric bud development (GO:0001657)
padj=0.073 ROBO2, FGFR2
regulation of osteoblast differentiation (GO:0045667)
padj=0.073 YAP1, ZHX3, FGFR2
regulation of cell−substrate junction assembly (GO:0090109)
padj=0.073 LIMCH1, EFNA5
Term
0.0 0.5 1.0 1.5 2.0
−log(Padj)

Figure 3.16. Putative functions of *LINC00499* in Astro L1-6 FGFR3 SLC14A1 investigated using two computational strategies. (A, B) GBA analysis for *LINC00449* in Astro L1-6 FGFR3 SLC14A1 showing the expression levels (A) and associated GO Biological processes (B) for the top 50 positively correlated (red) and top 50 negatively correlated genes (blue). (C, D) FasimTarget analysis for *LINC00499* showing the expression of the top 50 FasimTarget hits, ranked by the percentage of nuclei expressing them (C), and associated GO Biological processes (D). Dotplots are coloured according to the level of expression (scaled average normalised gene counts). The size of the dots indicates the proportion of nuclei expressing the gene per sample. Enrichment plots are coloured according to the MTG taxonomy[109].

Putative DNA interaction partners of *AC004852.2* included genes such as *CHL1* and *NOVA1* which were expressed in at least 70% of nuclei in the OPC L1-6 PDGFRA population (Fig 3.17C, Supp Table 3.32). *CHL1* and *NOVA1* were amongst the top 50 co-expressed genes showing a positive correlation with *AC004852.2,* and were the only genes shared between the GBA and LongTarget analyses (Fig 3.17A, Fig 3.17C, Supp Fig 3.30B). The top 50 FasimTarget hits (by percentage of nuclei expressing the gene) did not show a distinct corresponding or opposing expression pattern to that of *AC004852.2* (Fig 3.17C, Supp Table 3.32). They were enriched for terms such as proteoglycan metabolic process, negative regulation of cell differentiation, and positive regulation of vascular endothelial growth factor production (Fig 3.17D, Supp Table 3.32).

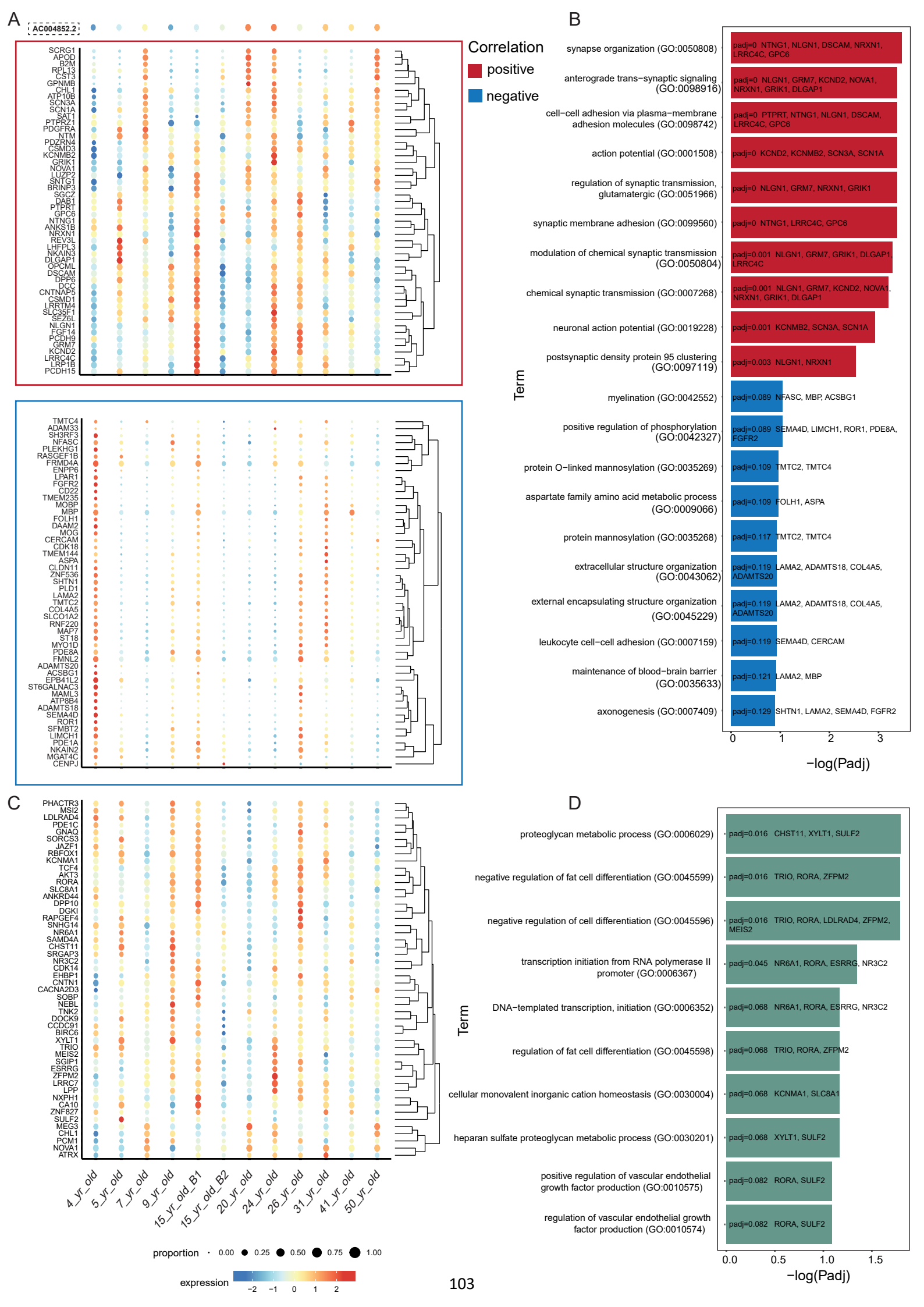

A
AC004852.2
SCRG1
APOD
B2M
RPL13
CST3
GPNMB
CHL1
ATP1B
SCN3A
SCN1A
SAT1
PTPRZ1
PDGFRA
NTM
PDZRN4
CSMD3
KCNMB2
GRIK1
NOVA1
LUZP2
SNTG1
BRINP3
SGCZ
DAB1
PTPRT
GPC6
NTNG1
ANKS1B
NRXN1
REV3L
LHFPL3
NKAIN3
DLGAP1
OPCML
DSCAM
DPP6
DCC
CNTNAP5
CSMD1
LRRTM4
SLC35F1
SEZ6L
NLGN1
FGF14
PCDH9
GRM7
KCND2
LRRC4C
LRP1B
PCDH15
TMTC4
ADAM33
SH3RF3
NFASC
PLEKHG1
RASGEF1B
FRMD4A
ENPP6
LPAR1
FGFR2
CD22
TMEM235
MOBP
MBP
FOLH1
DAAM2
MOG
CERCAM
CDK18
TMEM144
ASPA
CLDN11
ZNF536
SHTN1
PLD1
LAMA2
TMTC2
COL4A5
SLCO1A2
RNF220
MAP7
ST18
MYO1D
PDE8A
FMNL2
ADAMTS20
ACSBG1
EPB41L2
ST6GALNAC3
MAML3
ATP8B4
ADAMTS18
SEMA4D
ROR1
SFMBT2
LIMCH1
PDE1A
NKAIN2
MGAT4C
CENPJ
Correlation
positive
negative
B
synapse organization (GO:0050808)
anterograde trans−synaptic signaling (GO:0098916)
cell−cell adhesion via plasma−membrane adhesion molecules (GO:0098742)
action potential (GO:0001508)
regulation of synaptic transmission, glutamatergic (GO:0051966)
synaptic membrane adhesion (GO:0099560)
modulation of chemical synaptic transmission (GO:0050804)
chemical synaptic transmission (GO:0007268)
neuronal action potential (GO:0019228)
postsynaptic density protein 95 clustering (GO:0097119)
myelination (GO:0042552)
positive regulation of phosphorylation (GO:0042327)
protein O−linked mannosylation (GO:0035269)
aspartate family amino acid metabolic process (GO:0009066)
protein mannosylation (GO:0035268)
extracellular structure organization (GO:0043062)
external encapsulating structure organization (GO:0045229)
leukocyte cell−cell adhesion (GO:0007159)
maintenance of blood−brain barrier (GO:0035633)
axonogenesis (GO:0007409)
padj=0 NTNG1, NLGN1, DSCAM, NRXN1, LRRC4C, GPC6
padj=0 NLGN1, GRM7, KCND2, NOVA1, NRXN1, GRIK1, DLGAP1
padj=0 PTPRT, NTNG1, NLGN1, DSCAM, LRRC4C, GPC6
padj=0 KCND2, KCNMB2, SCN3A, SCN1A
padj=0 NLGN1, GRM7, NRXN1, GRIK1
padj=0 NTNG1, LRRC4C, GPC6
padj=0.001 NLGN1, GRM7, GRIK1, DLGAP1, LRRC4C
padj=0.001 NLGN1, GRM7, KCND2, NOVA1, NRXN1, GRIK1, DLGAP1
padj=0.001 KCNMB2, SCN3A, SCN1A
padj=0.003 NLGN1, NRXN1
padj=0.089 NFASC, MBP, ACSBG1
padj=0.089 SEMA4D, LIMCH1, ROR1, PDE8A, FGFR2
padj=0.109 TMTC2, TMTC4
padj=0.109 FOLH1, ASPA
padj=0.117 TMTC2, TMTC4
padj=0.119 LAMA2, ADAMTS18, COL4A5, ADAMTS20
padj=0.119 LAMA2, ADAMTS18, COL4A5, ADAMTS20
padj=0.119 SEMA4D, CERCAM
padj=0.121 LAMA2, MBP
padj=0.129 SHTN1, LAMA2, SEMA4D, FGFR2
Term
0
1
2
3
−log(Padj)
C
PHACTR3
MSI2
LDLRAD4
PDE1C
GNAQ
SORCS3
JAZF1
RBFOX1
KCNMA1
TCF4
AKT3
RORA
SLC8A1
ANKRD44
DPP10
DGKI
RAPGEF4
SNHG14
NR6A1
SAMD4A
CHST11
SRGAP3
NR3C2
CDK14
EHBP1
CNTN1
CACNA2D3
SOBP
NEBL
TNK2
DOCK9
CCDC91
BIRC6
XYLT1
TRIO
MEIS2
SGIP1
ESRRG
ZFPM2
LRRC7
LPP
NXPH1
CA10
ZNF827
SULF2
MEG3
CHL1
PCM1
NOVA1
ATRX
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
proportion
0.00
0.25
0.50
0.75
1.00
expression
−2
−1
0
1
2
D
proteoglycan metabolic process (GO:0006029)
negative regulation of fat cell differentiation (GO:0045599)
negative regulation of cell differentiation (GO:0045596)
transcription initiation from RNA polymerase II promoter (GO:0006367)
DNA−templated transcription, initiation (GO:0006352)
regulation of fat cell differentiation (GO:0045598)
cellular monovalent inorganic cation homeostasis (GO:0030004)
heparan sulfate proteoglycan metabolic process (GO:0030201)
positive regulation of vascular endothelial growth factor production (GO:0010575)
regulation of vascular endothelial growth factor production (GO:0010574)
padj=0.016 CHST11, XYLT1, SULF2
padj=0.016 TRIO, RORA, ZFPM2
padj=0.016 TRIO, RORA, LDLRAD4, ZFPM2, MEIS2
padj=0.045 NR6A1, RORA, ESRRG, NR3C2
padj=0.068 NR6A1, RORA, ESRRG, NR3C2
padj=0.068 TRIO, RORA, ZFPM2
padj=0.068 KCNMA1, SLC8A1
padj=0.068 XYLT1, SULF2
padj=0.082 RORA, SULF2
padj=0.082 RORA, SULF2
Term
0.0
0.5
1.0
1.5
−log(Padj)

Figure 3.17. Putative functions of *AC004852.2* in OPC L1-6 PDGFRA investigated using two computational strategies. (A, B) GBA analysis for *AC004852.2* in OPC L1-6 PDGFRA showing the expression levels (A) and associated GO Biological processes (B) for the top 50 positively correlated (red) and top 50 negatively correlated genes (blue). (C, D) FasimTarget analysis for *AC004852.2* showing the expression of the top 50 FasimTarget hits, ranked by the percentage of nuclei expressing them (C), and associated GO Biological processes (D). Dotplots are coloured according to the level of expression (scaled average normalised gene counts). The size of the dots indicates the proportion of nuclei expressing the gene per sample. Enrichment plots are coloured according to the MTG taxonomy[109].

3.10. Validation of NS-Forest markers using Visium spatial transcriptomics

To validate the cell-type-specificity of several of the NS-Forest markers, their expression levels were examined in Visium datasets that were generated, together with Ruvimbo Mishi (MSc student in the Hockman lab) for a subset of samples, namely the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old. The set of 23 snRNA-seq datasets was used as a reference to annotate the Visium datasets, thus allowing a direct comparison between the snRNA-seq and Visium cell types. A machine learning method (cell2location[178]) was used to estimate the level of expression of each gene in each cell type in the spatial transcriptomic datasets.

The expression of *LINC00499*, a marker for Astro L1-6 FGFR3 SLC14A1 in the snRNA-seq data, appeared to be unique to Astro L1-6 FGFR3 SLC14A1 across all four Visium samples (Fig 3.18A). The 15-year-old B1, 15-year-old B2, and 31-year-old samples showed a greater number of positive spots compared to the 4-year-old which only had one positive spot while the 31-year-old had the greatest number of positive spots (Fig 3.18A). Comparing *LINC00499* expression to the distribution of Astro L1-6 FGFR3 SLC14A1 across the tissue showed that *LINC00499* was expressed in a subset of the spots positive for Astro L1-6 FGFR3 SLC14A1 (Fig 3.18A-B). For example, in the 31-year-old sample, Astro L1-6 FGFR3 SLC14A1 was most abundant in the first cortical layer with several spots in this layer also being positive for *LINC00499* (Fig 3.18A-B). The cell-type specificity of expression was corroborated by the snRNA-seq data, which showed that *LINC00499* expression was most highly expressed in Astro L1-6 FGFR3 SLC14A1 (Fig 3.18C). Additionally, as expected from the differential expression analysis for this cell type, *LINC00499* was expressed at higher levels in the adult samples than the paediatric samples, with little expression observed in the 4-year-old and 5-year-old (Fig 3.18D).

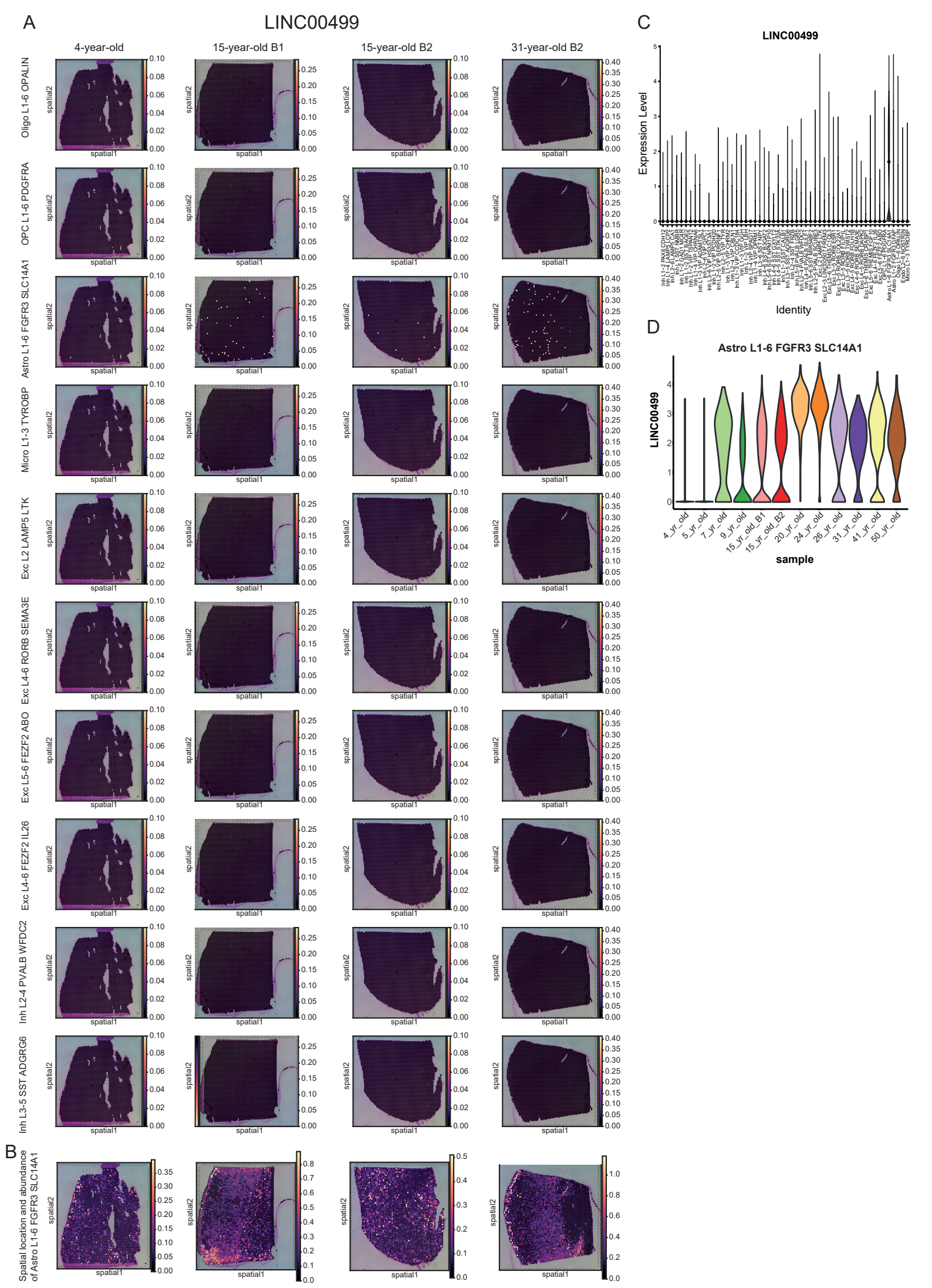

A
LINC00499
4-year-old
15-year-old B1
15-year-old B2
31-year-old B2
Oligo L1-6 OPALIN
OPC L1-6 PDGFRA
Astro L1-6 FGFR3 SLC14A1
Micro L1-3 TYROBP
Exc L2 LAMP5 LTK
Exc L4-6 RORB SEMA3E
Exc L5-6 FEZF2 ABO
Exc L4-6 FEZF2 IL26
Inh L2-4 PVALB WFDC2
Inh L3-5 SST ADGRG6
spatial1
spatial2
B
Spatial location and abundance of Astro L1-6 FGFR3 SLC14A1
C
LINC00499
Expression Level
Identity
D
Astro L1-6 FGFR3 SLC14A1
LINC00499
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
sample

Figure 3.18. Validation of *LINC00499* expression as a cell type-specific marker of Astro L1-6 FGFR3 SLC14A1 using Visium spatial transcriptomics. (A) Spatial plots showing the estimated expression level and location of *LINC00499* in a subset of cell types for the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old Visium datasets. The scale minimum was set to zero. For each sample, the 99.2% quantile of gene expression of each cell type was computed and the scale limited to the maximum value computed. (B) Spatial plot showing the estimated abundance and spatial location of Astro L1-6 FGFR3 SLC14A1 in each of the Visium datasets. The scale minimum was set to zero. The scale maximum was set to the 99.2% quantile of cell abundance of Astro L1-6 FGFR3 SLC14A1 for each sample. (C) Violin plot showing the level of expression and proportion of nuclei expressing *LINC00499* across all cell types in the snRNA-seq datasets. Error bars represent mean +-SEM. (D) Violin plots showing the level of expression and proportion of nuclei expressing *LINC00499* across all 12 samples within the Astro L1-6 FGFR3 SLC14A1 population.

The expression of *AC004852.2*, a marker of OPC L1-6 PDGFRA, was estimated to only localise to the equivalent OPC L1-6 PDGFRA population in the Visium data (Fig 3.19A). It appeared to be expressed in a moderate number of spots in the 31-year-old while being expressed in fewer than 10 spots in the 4-year-old, 15-year-old B1, and 15-year-old B2 (Fig 3.19A). Its spatial location corresponded with spots showing moderate abundance of OPC L1-6 PDGFRA cells (Fig 3.19B). In the snRNA-seq datasets, *AC004852.2* showed cell-type-specificity for OPC L1-6 PDGFRA with little off-target expression observed (Fig 3.19C). Within OPC L1-6 PDGFRA, *AC004852.2* was expressed at high levels across all 12 samples (Fig 3.19D).


AC004852.2
A
4-year-old
15-year-old B1
15-year-old B2
31-year-old B2
Oligo L1-6 OPALIN
OPC L1-6 PDGFRA
Astro L1-6 FGFR3 SLC14A1
Micro L1-3 TYROBP
Exc L2 LAMP5 LTK
Exc L4-6 RORB SEMA3E
Exc L5-6 FEZF2 ABO
Exc L4-6 FEZF2 IL26
Inh L2-4 PVALB WFDC2
Inh L3-5 SST ADGRG6
spatial1
spatial2
B
Spatial location and abundance of OPC L1-6 PDGFRA
C
AC004852.2
Expression Level
Identity
D
OPC L1-6 PDGFRA
AC004852.2
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
sample

Figure 3.19. Validation of *AC004852.2* expression as a cell type-specific marker of OPC L1-6 PDGFRA using Visium spatial transcriptomics. (A) Spatial plots showing the estimated expression level and location of *AC004852.2* in a subset of cell types for the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old Visium datasets. The scale minimum was set to zero. For each sample, the 99.2% quantile of gene expression of each cell type was computed and the scale limited to the maximum value computed. (B) Spatial plot showing the estimated abundance and spatial location of OPC L1-6 PDGFRA in each of the Visium datasets. The scale minimum was set to zero. The scale maximum was set to the 99.2% quantile of cell abundance of OPC L1-6 PDGFRA for each sample. (C) Violin plot showing the level of expression and proportion of nuclei expressing *AC004852.2* across all cell types in the snRNA-seq datasets. Error bars represent mean +-SEM. (D) Violin plots showing the level of expression and proportion of nuclei expressing *AC004852.2* across all 12 samples within the OPC L1-6 PDGFRA population.

The expression of *APBB1IP*, a marker of Micro L1-3 TYROBP, was largely specific to the microglial population in all four Visium datasets, however there appeared to be some expression in Exc L4-6 FEZF2 IL26, albeit at lower levels than in Micro L1-3 TYROBP (Fig 3.20A). The spatial location of spots expressing *APBB1IP* appeared to largely correspond with the estimated spatial location of the Micro L1-3 TYROBP cells which were distributed across the tissue sections of the four samples, with each sample having multiple positive *APBB1IP*-expressing spots (Fig 3.20B). *APBB1IP* showed high specificity for Micro L1-3 TYROBP in the snRNA-seq datasets (Fig 3.20C) and was expressed at high levels in this cell type across all the samples (Fig 3.20D). It was also expressed at low levels in Exc L4-6 FEZF2 IL26 in the snRNA-seq datasets – corroborating its expression in the Visium data (Fig 3.20C).


A
APBB1IP
4-year-old
15-year-old B1
15-year-old B2
31-year-old B2
Oligo L1-6 OPALIN
OPC L1-6 PDGFRA
Astro L1-6 FGFR3 SLC14A1
Micro L1-3 TYROBP
Exc L2 LAMP5 LTK
Exc L4-6 RORB SEMA3E
Exc L5-6 FEZF2 ABO
Exc L4-6 FEZF2 IL26
Inh L2-4 PVALB WFDC2
Inh L3-5 SST ADGRG6
spatial1
spatial2
B
Spatial location and abundance of Micro L1-3 TYROBP
C
APBB1IP
Expression Level
Identity
D
Micro L1-3 TYROBP
APBB1IP
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
sample

Figure 3.20. Validation of *APBB1IP* expression as a cell type-specific marker of Micro L1-3 TYROBP using Visium spatial transcriptomics. (A) Spatial plots showing the estimated expression level and location of *APBB1IP* in a subset of cell types for the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old Visium datasets. The scale minimum was set to zero. For each sample, the 99.2% quantile of gene expression of each cell type was computed and the scale limited to the maximum value computed. (B) Spatial plot showing the estimated abundance and spatial location of Micro L1-3 TYROBP in each of the Visium datasets. The scale minimum was set to zero. The scale maximum was set to the 99.2% quantile of cell abundance of Micro L1-3 TYROBP for each sample. (C) Violin plot showing the level of expression and proportion of nuclei expressing *APBB1IP* across all MTG cell types in the snRNA-seq datasets. Error bars represent mean +-SEM. (D) Violin plots showing the level of expression and proportion of nuclei expressing *APBB1IP* across all 12 samples within the Micro L1-3 TYROBP population.

SEMA3E, a marker of Exc L5-6 FEZF2 ABO for 10 of the samples from the snRNA-seq analysis, displayed high cell-type specific expression in the expected cell type across all four Visium samples (Fig 3.21A). A low level of *SEMA3E* expression was seen in several other cell types, including OPC L1-6 PDGFRA, Exc L4-6 FEZF2 IL26, and Exc L4-6 RORB SEMA3E (Fig 3.21A). The location of *SEMA3E* expression corresponded closely with the location of spots with the greatest estimated abundance of Exc L5-6 FEZF2 ABO (Fig 3.21B). In the snRNA-seq data, *SEMA3E* showed high expression in Exc L5-6 FEZF2 ABO with a median expression of zero in all other cell types (Fig 3.21C). Within Exc L5-6 FEZF2 ABO, it was expressed at high levels in almost all samples but showed no expression in the 24-year-old (Fig 3.21D).

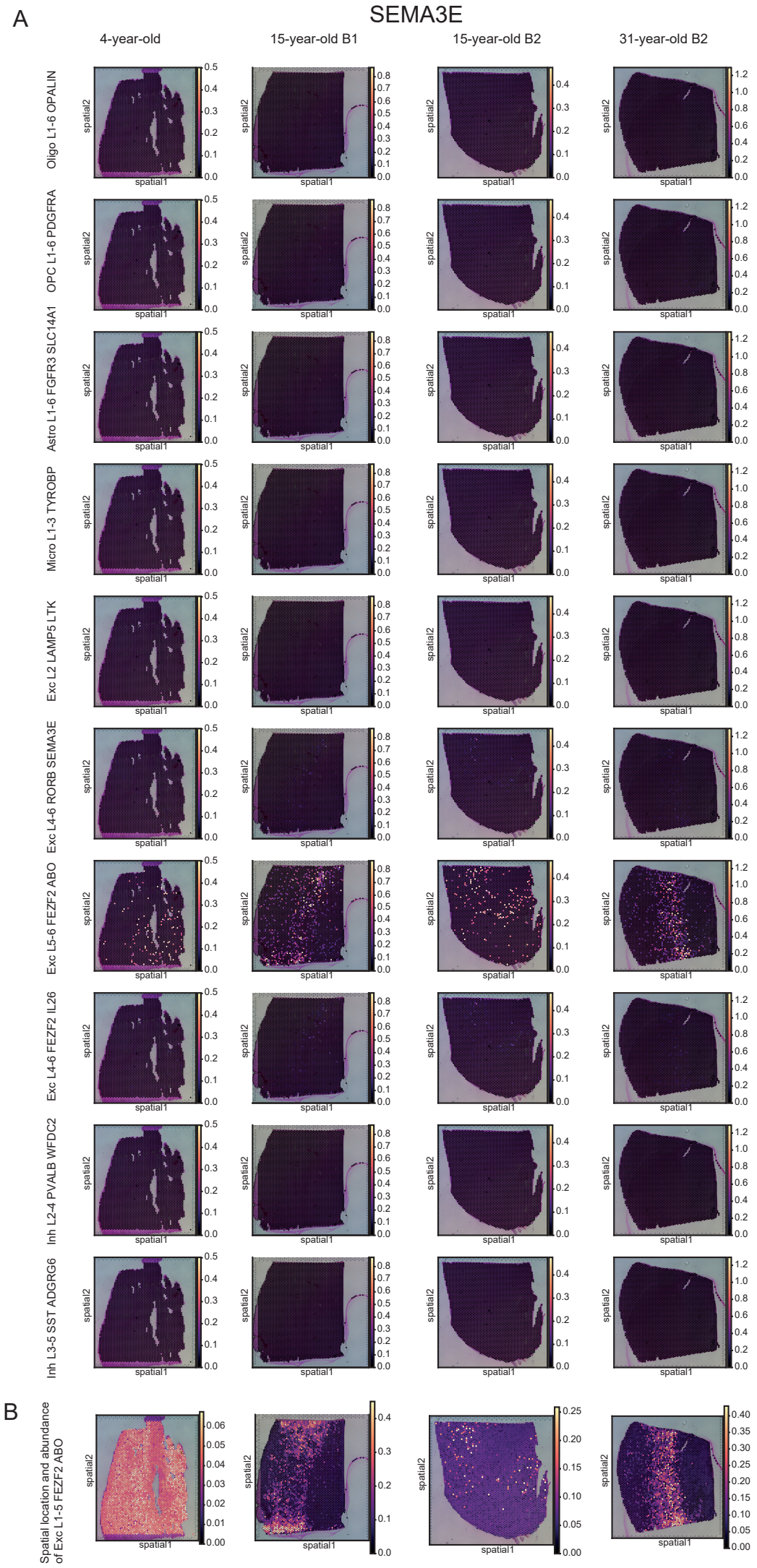

A
SEMA3E
4-year-old
15-year-old B1
15-year-old B2
31-year-old B2
Oligo L1-6 OPALIN
OPC L1-6 PDGFRA
Astro L1-6 FGFR3 SLC14A1
Micro L1-3 TYROBP
Exc L2 LAMP5 LTK
Exc L4-6 RORB SEMA3E
Exc L5-6 FEZF2 ABO
Exc L4-6 FEZF2 IL26
Inh L2-4 PVALB WFDC2
Inh L3-5 SST ADGRG6
spatial1
spatial2
B
Spatial location and abundance of Exc L1-5 FEZF2 ABO
C
SEMA3E
Expression Level
Identity
D
Exc L5-6 FEZF2 ABO
SEMA3E
4_yr_old
5_yr_old
7_yr_old
9_yr_old
15_yr_old_B1
15_yr_old_B2
20_yr_old
24_yr_old
26_yr_old
31_yr_old
41_yr_old
50_yr_old
sample

Figure 3.21. Validation of *SEMA3E* expression as a cell type-specific marker of Exc L5-6 FEZF2 ABO using Visium spatial transcriptomics. (A) Spatial plots showing the estimated expression level and location of *SEMA3E* in a subset of cell types for the 4-year-old, 15-year-old B1, 15-year-old B2, and 31-year-old Visium datasets. The scale minimum was set to zero. For each sample, the 99.2% quantile of gene expression of each cell type was computed and the scale limited to the maximum value computed. (B) Spatial plot showing the estimated abundance and spatial location of Exc L5-6 FEZF2 ABO in each of the datasets. The scale minimum was set to zero. The scale maximum was set to the 99.2% quantile of cell abundance of Exc L5-6 FEZF2 ABO for each sample. (C) Violin plot showing the level of expression and proportion of nuclei expressing *SEMA3E* across all MTG cell types in the snRNA-seq datasets. Error bars represent mean +-SEM. (D) Violin plots showing the level of expression and proportion of nuclei expressing *SEMA3E* across all 12 samples within the Exc L5-6 FEZF2 ABO population.

The 4-year-old and 15-year-old B2 samples appeared to be of lower quality than the 15-year-old B1 and 31-year-old samples based on the expression of previously described layer-specific markers[105] (Supp Fig 3.31). Most of the markers showed more distinct expression in the expected layers for the 15-year-old B1 and the 31-year-old compared to the 4-year-old and 15-year-old B2 (Supp Fig 3.31). Additionally, they appeared to be expressed in a greater number of spots in the 15-year-old B1 and the 31-year-old compared to the 4-year-old and 15-year-old B2 (Supp Fig 3.31). For example, *TRABD2A*, a layer V marker, showed clear expression in the expected location in the 15-year-old B1 and the 31-year-old (Supp Fig 3.31). However, in the 4-year-old dataset and 15-year-old B2 datasets its expression was more widely distributed and it was expressed in fewer spots than observed in the 15-year-old B1 and 31-year-old (Supp Fig 3.31).

In summary, to analyse the snRNA-seq datasets, stringent quality control measures were initially performed resulting in the removal of doublets and poor-quality nuclei. The datasets were integrated to align nuclei with similar transcriptomic profiles across the datasets and clustering analysis was performed to assign the nuclei to a group according to their expression profiles. The nuclei were then annotated into 7 broad cell types and 54 high resolution cell types using a combination of manual and automated methods. Subsequently, NS-Forest was used to identify small combinations of coding and non-coding marker genes defining each cell type for each sample. Three differential gene expression analysis tools were then applied to identify genes changing in their level of expression with age including DESeq2, Psupertime, and IDEAS. Additionally, proportion analysis was performed as a complementary approach to assess whether there was a change in the proportion of nuclei expressing the genes in each cell type over the course of brain maturation. Based on the above analyses, the putative functions of two lncRNAs of interest were investigated *in silico.* Furthermore, the expression of several genes was validated using Visium spatial transcriptomic datasets. Gene set enrichment analyses was performed on lists of genes at multiple points in the downstream analyses, revealing roles for the genes in various brain-related processes and neurological diseases.

Chapter 4: Discussion

The overall aim of this project was to address the gap in our understanding of changes in gene expression dynamics occurring during postnatal human brain maturation. As the state of the literature currently stands, there are no studies exploring this process at single-cell resolution, thus warranting the generation of paediatric snRNA-seq datasets using human brain tissue. To this end, this research involved the transcriptomic profiling and subsequent analysis of paediatric and adult cortical nuclei obtained from elective surgeries to treat epilepsy including datasets from 6 paediatric and 6 adult donors.

4.1. Data processing

Considering the 50-60% capture efficiency of the 10X Genomics platform[134], the expected number of nuclei targeted for each sample was 10 000 having loaded approximately 16 000 nuclei per reaction. However, the average number of barcodes recovered prior to filtering was 8390 (Inter Quartile Range (IQR): 6176- 9168) per sample which is slightly lower than expected and may be due to inaccurate concentration estimations following nuclei isolation. Nevertheless, for several samples including the 9-year-old T1, 9-year-old T2, and 15-year-old T1 B1, a greater number of barcodes were recovered than estimated (each > 13000 barcodes prior to filtering). Markedly, the mean number of reads per nucleus and the average sequencing saturation of the publicly available datasets were almost double that of the datasets generated in our laboratory. The sequencing saturation is a measure of the number of times the same UMI has been sequenced, which is affected by both the sequencing depth and the complexity of the cell types being sequenced[134]. A low sequencing saturation indicates a high probability of there being remaining unsequenced transcripts which can influence the ability to detect lowly expressed genes[134]. Interestingly however, a slightly lower number of genes were detected in the publicly available datasets compared to the datasets generated in our laboratory, despite the publicly available datasets being sequenced to a greater depth. This is likely due to the newer version of the single-cell chemistry platform used to generate our datasets (v3) which has been shown to promote the detection of a greater number of genes compared to the older chemistry (v2)[134], even when fewer reads are sequenced per cell[276]. Thus, despite the new datasets having a lower sequencing saturation than the publicly available datasets, the presence of a greater number of genes suggests that the quality of these datasets was sufficient to interrogate them further.

Single-cell data is notoriously noisy due to various technical artifacts which can complicate downstream interpretation[195,202]. For this reason, taking stringent quality control measures is advisable including the identification and removal of doublets[127,199]. Notably, the rate of doublets in droplet-based studies depends on the input concentration of cells or nuclei loaded[277]. For example, the estimated multiplet rate using the 10X Genomics platform is ~6.1% when 8000 barcodes are targeted compared to ~0.8% when 1000 barcodes are targeted[134]. Across the 23 datasets analysed in this study, the average doublet removal rate

was 12.62% (IQR: 11.65%-13.29%). Thus, the doublet rate is higher than the estimated rate for ~8000 targeted nuclei using the 10X Genomics platform which could be due to using a stringent doublet identification protocol comprising of three different computational tools. Whilst this approach may be over-sensitive in its doublet calling, it allows for greater confidence that the remaining nuclei following doublet removal are singlets.

Other quality control measures included cell and gene level filtering to ensure that poor quality nuclei were removed. After filtering, the remaining nuclei largely had similar numbers of genes expressed and transcripts per nucleus per sample. However, the 4-year-old and 26-year-old both had a greater median number of genes and UMIs detected, warranting appropriate normalization. Overall, the remaining nuclei showed uniformly high complexity (average log10Genes/UMI > 0.9) which is expected for brain cells since the majority of genes in the genome are expressed in brain tissue (high transcript diversity)[278]. On the other hand, the average mitoRatio per sample was very low (< 0.017). This was also expected since the experiment used nuclei (instead of cells) which do not express mitochondrial genes. The presence of some mitochondrial genes was likely due to some mitochondrial RNA from lysed cells attaching to the nuclear membrane after nuclei isolation[279]. While mitochondrial genes may represent an interesting source of variation in cells, their presence in the nuclear fraction is likely random and uninformative regarding the biology of the nuclei and so they were removed from downstream analysis.

Subsequent to quality control, differences in sequencing depth between nuclei were accounted for by normalizing the counts per nucleus. Additionally, the variance in expression levels across nuclei was scaled with the mean to ensure that it is not only highly expressed genes contributing to variation between nuclei[202]. In addition to removing mitochondrial genes from the gene matrix, the mitoRatio variable associated with each barcode was regressed out to mitigate any effect it may have on downstream clustering. The 23 datasets were then integrated in an attempt to remove technical differences between samples and align similar nuclei across the various datasets[205]. There appeared to be a clear batch effect on first observing that some nuclei separated away from other datasets in the cluster subsequently annotated as oligodendrocytes. However, closer inspection revealed that the four separating branches belonged to technical replicates, each within the same batch (9-year-old T1 and 9-year-old-T2 in batch D; 31-year-old T1 and 31-year-old T2 batch C). Additionally, these nuclei only separated away from the other datasets *after* not *prior* to integration suggesting that this observation represents an artefact of the integration process and is not indicative of a batch effect. Notably, the distance between data points on the UMAP depends on parameters such as the min_dist parameter from the RunUMAP function which tries to optimise the distances with regards to the local structure[212,280]. Additionally, visualising the data in two dimensions can make distances appear further than they would using more dimensions, and thus the distances between the points on the UMAP plot should not be interpreted as a direct measure of the extent of dissimilarity[212,280]. These parameters were not adjusted when generating the plots since the UMAP could still be annotated under the default parameters. Taken together, the outlying nuclei may not be

very different to the non-outlying nuclei but can cautiously be interpreted as biologically distinct. Since the oligodendrocyte population comprised the largest number of nuclei, it is likely the most diverse population and so it is possible that the outlying nuclei represent sub-types or a sub-state of oligodendrocytes.

4.2. Cell type annotation

To annotate the nuclei as one of the various cell populations in the brain, automated and manual annotation approaches were used. Manual annotation included generating violin plots to visualize the expression of previously described marker genes across the unannotated clusters[108–110]. However, this method was ambiguous as many markers showed off-target expression in multiple clusters and thus were not useful for clearly distinguishing cell types. Nevertheless, a label transfer approach was capable of assigning each nucleus into one of 54 cortical subtypes from the MTG reference dataset comprising 75 transcriptomically distinct cell types[109]. The annotations were validated by examining the expression of known cell type-specific marker genes in the various labelled populations. Reassuringly, the marker genes were expressed at higher levels in the expected cell types relative to other cell types, suggesting that the nuclei were largely annotated correctly.

The inability to identify all 75 cell types identified in the MTG reference dataset published by Hodge et al. (2019)[109] could be due to several reasons. Firstly, this could be due to a technical sampling bias. Hodge et al. (2019) specifically dissected each cortical layer from MTG sections and enriched for nuclei from cortical layers with higher cell type diversity[109]. In this study no layer-specific dissection or enrichment was performed which may have led to under sampling the diversity of cell types. Indeed, seven cell subtypes had fewer than 50 nuclei across all 23 samples suggesting possible scarcity of certain cell subtypes. Secondly, Hodge et al. (2019) used a combination of known cell type-specific markers and layer information to annotate the nuclei. This could not be done in this study due to the lack of information regarding which layer the nuclei originated from. Instead, the label transfer method was used, which is a predictive method that scores each nucleus according to its similarity with each reference cell type[205]. This is based on correlating the expression of each nucleus (query) to anchors comprising of pairs of reference and query nuclei[205]. It is important to note that label transfer methods can be error prone due to incorrect pre-processing of query or reference datasets[281]. They are also intrinsically biased in that they force each nucleus to take on the annotation with the maximum predictive score without using a threshold for dissimilarity (i.e all nuclei are annotated even if they show low predictive scores for all cell types)[205]. Lastly, Hodge et al. (2019) used the Smart-seq platform to generate their snRNA-seq data, which usually results in a greater sequencing depth per nucleus[282], whereas this study used the 10X Genomics platform[134] and did not saturate sequencing depth. This difference may have precluded the identification of some cell subtypes that require a larger number of transcripts to be profiled in order to capture their molecular diversity.

One approach to validate the cell type annotations is to conduct a similarity analysis comparing the query cell types against the reference cell types[109]. Based on the method

used here, the non-neuronal populations in the query dataset showed exclusive correlations to their corresponding cell types in the reference dataset whereas the excitatory and inhibitory neuron subtypes appeared to correlate with multiple other excitatory and inhibitory neuron subtypes respectively. Although this may be due to the non-neuronal subtypes having more robust and distinctive transcriptomic profiles, it is possible that they would also correlate with multiple subtypes had they been annotated to the equivalently high resolution of the neuronal subtypes. In line with this hypothesis, a similarity analysis conducted by Hodge et al. (2019)[109] comparing the middle temporal gyrus dataset to the prefrontal cortex dataset from Habib et al. (2017)[256] found that the oligodendrocyte and astrocyte populations in the query dataset mapped to multiple oligodendrocyte and astrocyte subtypes respectively in the reference datasets. The low specificity observed when correlating query and reference cell subtypes at high resolution may also be due to the method of computing similarity not being sophisticated enough to robustly distinguish between different subtypes of cells. Both the similarity analysis used in this study and that described above in Hodge et al. (2019)[109] make use of correlating query and reference datasets using a subset of genes based on their beta score. The beta score threshold selects for genes that have a high classification power in terms of a relative expression level difference between a target cell type and other off target cell types[219]. However, these genes may still be expressed in off-target cell types and so using these genes to correlate various cell types may not sufficiently resolve differences between cell types[219]. An alternative and potentially more robust method of mapping query to reference cell types is described by Johansen et al (2019)[283]. This approach uses an unsupervised deep learning integration method to compute the proportion of query and reference nuclei that co-cluster for each pair of query and reference cell types. Hodge et al. (2019) applied this method to compute the cell type homology between mouse and human cortical cell types[109] but in theory this approach could also be used in future to validate the label transfer method used in this study.

To determine whether there is a change in the abundance of cortical cell types with age, the proportion of nuclei per cell type per sample were plotted. This revealed an apparent increase in the proportion of oligodendrocytes during the early postnatal period (ages 4 to 9). However, this interpretation is limited by the small sample size which makes this finding non generalizable. Additionally, the relative abundance of various cell types likely depends on the size of the tissue sampled, which was variable, and on the random nature of sampling. Moreover, this method does not account for the fact that changes in the proportion of one cell type can alter the proportion of another cell type[284]. To try and minimise this effect, the abundance of nuclei in each subtype can be scaled to the total number of nuclei in the broad cell type instead of to all nuclei[116]. Alternatively, a more suitable strategy for estimating changes in cell type abundance between groups may be to perform compositional data analysis using methods such as the recently published Cacoa method, which includes the estimateCellLoadings() function[285]. This approach may be more robust as it accounts for the fact that changes in the proportion of one cell type can affect the proportion of another. Further validation of this approach could include histological

analysis or analysis of spatial transcriptomic datasets[284] which would allow for the control of the size and the region of the tissue analysed. In addition to plotting the proportion of nuclei per cell type for each sample, this parameter was also analysed for each technical replicate to assess the variability between tissue pieces from the same sample. While the cell composition of the replicates was strikingly similar for most samples, the replicates of the 5-year-old and 7-year-old had disparate cell compositions. This is likely due to the region of origin of the tissue piece used to generate the datasets. For example, the oligodendrocyte population appeared to be more abundant in the 7-year-old T1 replicate compared to the T2 replicate. This may be a result of the T1 tissue having more white matter than the T2 tissue, since oligodendrocytes typically reside in the white matter and deeper cortical layers[286].

With regards to assessing the QC metrics of the annotated cell types, the neuronal cell types were observed to have a larger number of genes expressed per nucleus than non-neuronal cell types which is to be expected considering the greater complexity of neuronal cells, and is in agreement with the findings from previous studies using either the 10X Genomics or Smart-seq 4 method[108,109]. Overall, the median number of genes detected per nucleus was approximately 2500 genes for each cell type which is considerably lower than the approximate 7500 genes detected for the cell types in Hodge et al. (2019)[109]. This may be due to Hodge et al. (2019) using the Smart-seq V4[282] plate-based platform which likely resulted in a greater transcript capture efficiency per nucleus than the 10X Genomics droplet-based method in which the reagents are shared across multiple nuclei[134]. Indeed, comparing the plate-based and droplet-based methods applied to human, marmoset, and mouse motor cortex tissue in Bakken et al. (2021)[108], it appears that a greater number of genes were detected using Smart-seq V4 compared to the 10X Genomics platform. Similar to Hodge et al. (2019) and Bakken et al. (2021), the excitatory neuron subtypes in this study had a greater number of genes detected per nucleus than the inhibitory neuron subtypes with Exc L2−3 LINC00507 FREM3 and Exc L4−5 RORB FOLH1B among the cell types with the highest median gene detection. Markedly, the Exc L2−3 LINC00507 FREM3 population has previously been shown to have high transcriptomic diversity between nuclei and has itself been split into multiple subtypes, which may account for the larger transcript diversity within this cell-type[109] .

4.3. NS-Forest marker gene analysis

In order to explore the diversity of the brain cell atlases, marker gene analysis can be performed to identify genes that define each of the annotated cell types. The NS-Forest tool is different to other tools of its kind in that it tries to identify the smallest combination of marker genes required to distinguish cell types allowing for a scalable principle by which to define cell types[219]. Previously, Hodge et al. (2019)[109] and Aevermann et al. (2021)[219] applied the NS-Forest method to their MTG dataset (used as the reference for annotating our dataset), making these relevant studies to compare to this analysis. They applied the NS-Forest method to their merged dataset, comprising post-mortem and surgically resected temporal cortex samples from 8 individuals aged 24-66. Where Hodge et al. (2019) used NS-

Forest v1.3, Aevermann et al. (2021) used NS-Forest v2.0, with the main difference being that v2.0 was modified to filter out negative markers and prioritise binary markers (which have unique expression in one cell type) over markers that have off-target expression.

The average number of marker genes required per sample in our study to distinguish up to 54 cell types was 122.5 which equates to 2.4 markers required per cell type. Similarly, Aevermann et al. (2021) and Hodge et al. (2019) required 157 and 155 marker genes respectively to distinguish the 75 different cell types which corresponds to an average of 2.3 (Aevermann) and 2.4 (Hodge) markers per cell type. Plotting the expression of the marker genes against the corresponding cell types revealed a distinct diagonal line with the off-diagonal showing low relative expression to the diagonal, indicating global binary expression of the marker genes. Notably however, the NS-Forest markers for Micro L1-3 TYROBP, including *APBB1IP, ADAM28, DOCK8*, and *C3,* which are all known markers of microglia, were also expressed at low levels in Exc L4-6 FEZF2 IL26. While the similarity analysis suggested that these populations are transcriptomically distinct, the expression of microglial markers in the Exc L4-6 FEZF2 IL26 suggests that the label transfer method may not have accurately annotated this population and that at least a subset of these nuclei may belong to the microglia population.

To assess the usefulness of the marker genes in classifying cell types, two metrics, the F-beta score and the binary expression score were computed. The F-beta score is computed for a combination of marker genes and is a measure of the classification power of the genes based on their relative expression difference between the target and off-target clusters[219]. The binary expression score on the other hand is computed for individual genes and measures the uniqueness of a gene's expression in the target cluster compared to off-target clusters[219]. This is computed by assessing each gene's absolute level of expression and the proportion of nuclei expressing it in the target cluster compared to off-target clusters. Comparing the average F-beta score for each study showed a score of 0.67 (this study), 0.68 (Aevermann et al. (2021)), and 0.71 (Hodge et al. (2019)), whereas the average binary expression scores for each of the studies was 0.96 (this study), 0.94 (Aevermann et al. (2021)), and 0.72 (Hodge et al. (2019)). Thus, although there is a small decrease in the classification power of our markers compared to Hodge et al. (2019)[109], the increase in their binary expression scores makes them more useful for downstream applications such as histological or spatial transcriptomic validation[219]. Interestingly, the median average F-measure score for the marker gene combinations was slightly lower in the paediatric samples than adults suggesting that the classification power of the marker genes improves as the brain matures.

The minimal marker genes identified in this study showed greater overlap with those identified in Aevermann et al. (2021) than those from Hodge et al. (2019) which was expected since I used the same updated version of the NS-Forest tool as Aevermann et al. (2021)[219]. Interestingly, *STK32A* was shared between all three studies as a marker for Inh L4-5 SST STK32A. This gene encodes a kinase which has recently been characterised as having specificity for serine, threonine, and tyrosine residues[287] and was identified as a cell type-

specific marker of alpha motor neurons in mice[288]. Additionally, it was found to be upregulated in primary cultures of rat hippocampal neurons in response to Wnt signaling[289], a pathway important for regulating various neurodevelopmental events including neurogenesis and synaptic plasticity[290,291].

By applying NS-Forest to each sample individually, I was able to expose both the commonalties and differences between samples. Notably, there appears to be high inter-individual variation between samples with the majority of markers being specific to a single sample. Additionally, our markers showed low overall consensus with Aevermann et al[219] and Hodge et al[109]. The most plausible explanation for this is that there was greater diversity in our data since I applied the NS-Forest method to each sample individually where Aevermann et al. (2021) and Hodge et al. (2019) applied it to their merged datasets, in effect treating them as a single sample. Alternatively, it is possible that the low consensus is due to incorrect annotation of our datasets such that the cell types annotated in our datasets do not correspond to those annotated in the MTG dataset. However, validation of our annotations by directly comparing the similarity of our cell types to the MTG cell types suggests that the annotations are accurate. Moreover, 34 markers did overlap with Aevermann et al. (2021) and Hodge et al. (2019) across 30 diverse neuronal and nonneuronal cell types, suggesting that our annotations are correct.

Amongst the markers that overlapped multiple samples was *SEMA3E*, which was shared between 10 samples as a marker for Exc L5-6 FEZF2 ABO and has not previously been described as a marker for this cell type. Curiously however, *SEMA3E* was identified as a marker for the Exc L4-6 RORB SEMA3E cell type in Hodge et al. (2019)[109] raising the possibility that Exc L5-6 FEZF2 ABO in our dataset more closely resembles Exc L4-6 RORB SEMA3E than Exc L5-6 FEZF2 ABO from the MTG dataset. To exclude this possibility, the similarity analysis showed a slightly greater correlation between the Exc L5-6 FEZF2 ABO populations across the two studies than between Exc L5-6 FEZF2 ABO and Exc L4-6 RORB SEMA3E. Additionally, analysis of *SEMA3E* expression in the Visium datasets revealed that it is more strongly expressed in Exc L5-6 FEZF2 ABO than in Exc L4-6 RORB SEMA3E. Another explanation for the discrepancy could be that Hodge et al. (2019) and Aevermann et al. (2021) applied the analysis to merged adult datasets (the MTG reference dataset) instead of to individual adult and paediatric datasets. This can be tested in a future study by applying NS-Forest to each of the individual datasets comprising the MTG reference dataset and including additional paediatric snRNA-seq datasets. Notably, the *SEMA3E* gene encodes a protein with a semaphorin domain which functions through its receptor, *PLXND1*, to regulate axon guidance during pre and post-natal neural circuit development[292,293]. In support of its role in establishing neuronal circuits, the majority of the 10 datasets which had *SEMA3E* as a marker were paediatric samples. Nonetheless, since it remains a marker throughout maturation, it likely continues to play a key role in neuronal function even after the neural circuits have matured. In agreement with this, there is evidence that *SEMA3E* may function in regulating cell adhesion and the formation of synapses in a subset of granule cells in the

adult hippocampus whilst negatively regulating the number of stem cell niches in the subgranular zone[292].

In addition to markers shared between multiple samples, markers were also identified which were shared between paediatric samples only or shared between adult samples only. This included the neuromodulator, *PTH2R,* which was a marker for Inh L1-3 VIP GGH in paediatric samples. Notably, this gene encodes a G protein coupled receptor which binds *PTH2* to regulate various behavioural responses including having anti-anxiolytic and anti-depressive effects[294]. An important clarification is that the apparent specificity of these minimal markers for paediatric or adult cell types does not necessarily correspond to a difference in the *level* of expression of these markers between paediatric and adult tissue. Instead, these markers were determined as optimal minimal markers by NS-Forest based on their classification power and binary expression compared to other marker genes[219]. Nonetheless, there were several paediatric- or adult-specific markers that also showed a change in their level of expression. For example, in the NS-Forest analysis, *LINC00499* was largely shared between adult and not paediatric samples in Astro L1-6 FGFR3 SLC14A1 and was also upregulated in adults in this cell type based on differential expression analysis. Future studies should validate whether the paediatric and adult-specific minimal markers identified in this study are capable of distinguishing these two groups using a larger sample size.

Several subtypes including Exc L3-5 RORB FILIP1L, Exc L4-5 RORB FOLH1B and Exc L3-5 RORB TWIST2 had a larger number of non-coding than coding minimal markers whereas the non-neuronal subtypes all had a larger number of protein coding genes than non-coding genes serving as minimal markers. This could be indicative of non-coding genes contributing to the diversity of neuronal subtypes which is supported by previous studies showing that non-coding genes are highly cell type specific and may be better markers of mouse neuronal cell types than protein coding genes[129]. Both this study and Aevermann et al. (2021) highlight the relevance of non-coding genes as markers of cortical cell types with 41.8% of the markers in our study being non-coding genes compared to 24% in Aevermann et al. (2021)[219]. The means by which noncoding genes contribute to cell type diversity and complexity remains to be fully established[219]. Nevertheless, there is evidence that they may achieve this through various mechanisms, including altering the repertoire of available proteins through the regulation of alternative splicing[295].

Overall, applying NS-Forest to each dataset individually allowed me to assess the generalisability of the current MTG cell atlas annotations to individual samples[109]. Based on the similarity analysis, the annotations of our dataset appears to be correct indicating that the cell atlas is applicable to diverse samples. However, the *de novo* marker analysis revealed individual variability within these cell types, which may be relevant to understanding how patients differ from each other. In turn, this may lend insight into the aging process or disease state. Importantly, the NS-Forest tool is a classification strategy and thus differences in marker genes between individuals for a given cell type may only be differences at a definitional level which remain to be verified at a functional level.

4.4. Assessing changes in gene expression levels as the brain matures

To identify genes showing a change in their expression dynamics with age, three different tools were used, namely DESeq2's LRT method[209], PSupertime[226], and IDEAS[222] which were all applied to each cell type individually. PSupertime and DESeq2 could be directly compared since they were both designed to examine changes between various epochs of brain maturation whereas IDEAS was only intended to identify differences between the broad paediatric and adult groups. For both the DESeq2 and Psupertime analyses, there was a positive correlation between the number of analysed nuclei and the number of DEGs identified per cell type. The number of DEGs identified with IDEAS showed less dependency on the number of nuclei and in fact showed a weak negative correlation between these parameters. In agreement with this observation, pseudobulk methods such as DESeq2 have been shown to have enhanced sensitivity with an increase in the number of cells sampled[296]. Notably, IDEAS estimates the distribution of a given gene's expression across all nuclei of the cell type of interest for each sample and then assesses whether the sample-specific distributions are significantly different between the conditions that are being compared (i.e. paediatric and adult samples)[222]. Thus, one possibility to explain the inverse relationship between number of nuclei and number of DEGs identified by IDEAS is that populations with large numbers of nuclei are more heterogeneous and consequently there is reduced power to detect differences between groups[297]. This reduction in power may be more pronounced using IDEAS compared to DESeq2 or Psupertime due to IDEAS assessing differences in both the mean and variance between groups, whereas DESeq2 and Psupertime are only designed to assess differences in the mean[222]. To overcome this effect in the IDEAS analysis, large cell populations could be further split into subtypes as this may reduce the overall heterogeneity in the population thereby promoting the identification of more DEGs[297].

An exploration of the intersection of the DEGs determined for each cell type for each of the analysis methods separately revealed numerous DEGs to be cell type-specific even amongst closely related cell populations. In particular, more than half of the DEGs identified by DESeq2 were unique to a single cell type. This finding supports the hypothesis that resolving different cell types is important in the analysis of age-dependent transcriptomic changes in the brain. With regards to the coding status of the identified DEGs, the vast majority were protein-coding genes. This may be because non-coding genes are generally expressed at lower levels[298], and there is evidence that most DE tools have lower predictive power for lowly expressed genes compared to moderately or highly expressed genes[186]. Alternatively, it is possible that this is not a false negative finding and that most temporally regulated genes during brain maturation are protein-coding genes. In this case, we may speculate that generally, non-coding genes could be serving essential regulatory roles in specific cell types and are thus expressed at comparable levels across age groups, similar to house-keeping genes. The results of the NS-Forest analysis lend support to this hypothesis as many of the minimal marker genes identified were non-coding genes and were shared across multiple paediatric and adult samples suggesting that they may carry out fundamental functions in those cell types.

A recent single-cell analysis comparing temporal cortex samples from 75 individuals aged 18-83 years old was able to refine the expression of temporally regulated genes identified from bulk studies to specific cell types, including refining the age-dependent expression of *ZBTB16* to oligodendrocytes[116]. The current study now expands on this to include data from the paediatric population. Indeed, gene set enrichment analysis on the various lists of DEGs revealed numerous examples of genes that have previously been described in the GTEx Aging Signatures database of genes that were either up or downregulated in the brain with age according to bulk RNA-seq analysis[299]. For example, *ROBO2* has previously been found to decrease in expression in the brain between age 20 and 50 according to the GTEx Aging Signatures database[299] and was identified in this study as being downregulated in the adult datasets compared to the paediatric datasets in Exc L2 LAMP5 LTK and Exc L3-5 RORB ESR1 in both the IDEAS and PSupertime analyses. While the function of *ROBO2* as an axon guidance molecule during pre-natal development has been well characterized[300], its function during post-natal brain maturation had not been characterized until recently, despite evidence of its continued expression in the post-natal period[301]. To address this gap, Blockus et al. (2021)[301] used a *Robo2* conditional knockout mouse model to reveal a role for *ROBO2* in synapse formation in a subset of hippocampal excitatory CA1 pyramidal neurons. Based on their investigation, they concluded that dual functions of molecules such as *ROBO2* may help to explain the integration of pre-natal developmental processes, such as axon guidance, with post-natal developmental processes such as synaptogenesis[301]. Additionally, they proposed that *ROBO2's* expression in hippocampal excitatory CA1 pyramidal neurons may facilitate the establishment of synapse specificity whereby individual CA1 pyramidal neurons form synaptic connections with other excitatory neurons instead of inhibitory neurons – a process required for proper circuit formation[302]. Our study corroborates this by showing specific temporal regulation of *ROBO2* in subtypes of excitatory neurons over the post-natal maturation period which aligns with the timing of peak synaptogenesis[50,51]. In contrast to the *ROBO2* expression trajectory, the lncRNA, *LINC00499,* increased in expression with age in Astro L1-6 FGFR3 SLC14A1 based on DESeq2 and Psupertime. This agrees with previous findings from the GTEx database describing its upregulation in the brain with age, and refines *LINC00499's* temporal regulation to a specific cell type.

4.4.1 DESeq2 and Psupertime consensus analysis

DESeq2 and Psupertime were both applied to identify genes which show a change in their level of expression as a function of time, using independent strategies. Where DESeq2 uses a pseudo bulking strategy that aggregates counts across nuclei for each sample[209], Psupertime is based on pseudotime trajectory inference methods and explores gene expression changes as a trajectory across individual nuclei[226]. Pseudobulk differential expression analysis methods, such as DESeq2, have been found to outperform single-cell methods in terms of their sensitivity and specificity[186,296], although this was not compared to pseudotime methods such as Psupertime. On the other hand, Psupertime is specifically designed for time-series sc/snRNA-seq data and is superior to other pseudotime methods as it does not assume that age represents the greatest source of variation in the data[226].

Thus, it is useful for identifying DEGs when the changes in the level of expression are subtle or there are additional sources of variation in the data[226]. Notably, DESeq2 and Psupertime are not designed to identify genes which show a difference in variance between groups. Additionally, a shortfall of Psupertime is that it does not incorporate covariates in its design formula[226] which may exacerbate false positives by not accounting for the effect of potential confounders. Identifying DEGs common to both tools leverages the advantages of each tool while reducing the effect of their shortcomings. This is useful for identifying DEGs with high confidence which may be important for postnatal human brain maturation.

To this end, consensus genes were identified between DESeq2 and Psupertime which included several interesting patterns of genes that were either up- or downregulated over the course of brain maturation. In the Exc L2-3 LINC00507 FREM3 population, the set of consensus upregulated genes included *LINC-PINT*. This gene has previously been found to be upregulated in the human brain with age in a DNA microarray study using more than 1000 post-mortem specimens spanning embryonic development, childhood, and adulthood[143]. Additionally, the aforementioned single-cell study by Johansen et al (2022) of 75 adult temporal cortex samples also determined a change in the expression levels of *LINC-PINT* with age and refined this to excitatory neuron subclasses[116]. Here I add to this knowledge base by including information regarding the cell type-specific expression of *LINC-PINT* in the paediatric population relative to the adult population. Although the exact function of this gene in neurons is yet to be characterized, there is evidence that it may function as a transcriptional silencer of genes that promote neurodegeneration by interacting with the polycomb repressor complex[303,304], and thus its upregulation in excitatory neurons with age may be required for healthy brain aging. Recently, an alternatively spliced isoform of *LINC-PINT* has been shown to form a circular RNA (circRNA) encoding a short peptide implicated in glioblastoma[305]. Both the levels of *LINC-PINT* circRNA and its peptide have been shown to have reduced expression in glioblastoma samples compared to normal human brain tissue[305]. Moreover, the overexpression of the encoded peptide reduced glioblastoma cell proliferation *in vivo* and *in vitro*[305]. Importantly, the incidence of glioblastoma increases with age[306], warranting further interrogation of temporally regulated molecules such as *LINC-PINT* which may help to explain the increased risk with age.

In addition to *LINC-PINT*, *BAIAP3* was another gene upregulated with age in Exc L2-3 LINC00507 FREM3. It has been implicated in synaptic functions such as neurotransmitter release by influencing the SNARE-dependent anterograde trafficking pathway[307]. Additionally, it has been implicated in major depressive disorder by serving as a positive regulator of serotonin release from synapses[308]. Amongst the genes downregulated with age in Exc L2-3 LINC00507 FREM3 were *MYO16* and *FNBP1L,* which both have actin-related functions[309,310]. Markedly, increased levels of *MYO16* have been observed in the brains of patients with Schizophrenia compared to controls[311], implicating Exc L2-3 LINC00507 FREM3 in the pathogenesis of Schizophrenia. In agreement with this, dysregulated transcriptional control in upper layer cortical neurons, including layer 2 and 3 excitatory neurons have previously been implicated in Schizophrenia[284].

FNBP1L was also found to decrease in expression with age in Exc L2 LAMP5 LTK and is a known regulator of neurite outgrowth[312]. Additionally, it has been associated with both childhood and adult intelligence in GWAS studies[313,314] and has recently been discovered as a rare variant in Alzheimer's disease (AD) risk, with excitatory pyramidal CA1 neurons implicated in the pathology[315]. In parallel to this, *APOE*, which is the strongest genetic risk factor for AD[316], was also downregulated with age in Exc L2 LAMP5 LTK. GSEA revealed that both *FNBP1L* and *APOE* were associated with receptor-mediated endocytosis – a process essential for synaptic plasticity[317]. Taken together, it is possible that the dysregulation of these two genes within specific excitatory neuron subtypes during brain maturation may increase risk for AD due to disruption of synaptic plasticity, which is a well-characterised early event in AD pathology[318]. Indeed, there is evidence supporting the upregulation of neuronal ApoE expression as an early hallmark in the clinical progression of AD[319].

In Exc L3-5 RORB ESR1, the lncRNA, *STXBP5−AS1*, was found to increase in expression over the course of human brain maturation. Intriguingly, variants of this gene have been associated with Attention-deficit/hyperactivity disorder (ADHD) in both children and adults with a greater association of this gene observed in adult ADHD than in childhood ADHD[320]. The normal functioning of the lncRNA is proposed to act via the SNARE complex and increased levels of its expression have been negatively correlated with impulsivity. Markedly, variants in the SNARE complex have been proposed to play distinct roles at different developmental periods, including exerting age-specific effects on ADHD behaviours[321]. Our study corroborates this by showing that *STXBP5-AS1,* which indirectly regulates SNARE complex formation[320], changes in its expression level with age in Exc L3-5 RORB ESR1 which may help to explain the age-specific effects of the SNARE complex in various psychiatric conditions[321].

Curiously, *ARL17B* was differentially expressed in 27 different cell types according to the DESeq2 analysis and showed overlap between the DESeq2 and Psupertime analyses in OPC L1-6 PDGFRA, Astro L1-6 FGFR3 SLC14A1, and Exc L2 LAMP5 LTK, where it showed higher expression in epoch 5 compared to the other epochs. While the function of *ARL17B* has not been characterised, it belongs to a family of G proteins thought to play a role in membrane trafficking and cytoskeleton organisation[322]. Its expression has previously been observed in excitatory neurons, inhibitory neurons, oligodendrocytes, OPCs, astrocytes, and microglia[108]. Additionally, an AD-associated genomic variant has been associated with an increase in *ARL17B* expression in all major brain cell types[323]. Furthermore, variants associated with progressive supranuclear palsy have been associated with increased *ARL17B* expression in whole tissue brain samples comprising multiple different cell types[324]. Thus, while single-cell analysis can reveal cell type-specific expression, it is also useful for exposing genes that show synergistic changes in expression levels across different cell types.

CDH20 was one of the consensus genes upregulated with age in Oligo L1-6 OPALIN. This gene has previously been found to be expressed in a subtype of oligodendrocytes referred to as Olig1 in a snRNA-seq study of white matter from post mortem adult brain tissue[325].

Based on the expression pattern observed here it is possible that the Olig1 subtype may be more abundant in adult brains than in children. *GREB1L* also increased in expression with age in Oligo L1-6 OPALIN and has previously been found to be upregulated in astrocytes with age[326] – a signature which can now be modified to include oligodendrocytes. Furthermore, *QDPR* and *SNCA* were also upregulated in oligodendrocytes with age, and both play a role in dopamine biosynthesis which is a function that has not previously been described in oligodendrocytes. Considering that the dopaminergic system has been implicated in several neurological conditions[327] and that oligodendrocytes have also been implicated in various neuropsychiatric conditions[328], the alteration in expression of *SNCA* and *QDPR* in Oligo L1-6 OPALIN during brain maturation warrants further investigation. In contrast to *CDH20*, *QDPR,* and *SNCA, SGCZ* displayed a decreasing expression profile with age in Oligo L1-6 OPALIN with particularly high expression in the first epoch relative to other epochs. Although *SGCZ* expression has been characterised in neurons and OPCs[108], this is the first study to report its expression in oligodendrocytes which is likely due to previous single cell studies being carried out on adult samples in which it is expressed at very low levels.

Amongst the consensus genes which were downregulated in astrocytes during brain maturation was *ADORA2B,* which has previously been found to be upregulated in astrocytes in the early post-natal period in mouse brains. There is evidence that its expression in astrocytes during this period functions to terminate excessive synaptogenesis by negatively regulating the astrocytic expression of mGluR5 which promotes synaptogenesis[329]. Our study supports a similar role in humans by confirming the higher expression of *ADORA2B* in astrocytes during childhood.

Similar to the expression trajectory of *ADOR2B*, members of the small GTPase superfamily, *RHEB* and *RAB8B,* were highly expressed in epoch1 and subsequently downregulated in Micro L1-3 TYROBP. *RHEB* is a well-characterised regulator of mTOR signalling which is fundamental to eukaryotic cellular metabolism[330,331]. Its function was initially characterised in the brain where it was shown to have activity-dependent expression in neurons and is therefore proposed to play a role in long term potentiation[332]. Notably, its expression profile was shown to slightly decrease in the rat brain between the early postnatal period and adulthood based on Northern blot analysis[332]. Moreover, knockout of *Rheb* in adult and old-aged mice resulted in a reduced inflammatory response by microglia to lipopolysaccharide challenge, an effect which was more pronounced in the older mice than the younger mice[333]. While post-natal mice were not used in the above-mentioned study[333], we can infer that the upregulation of *RHEB* in microglia during the post-natal period, observed in our data, would likely enhance the microglial response to inflammatory challenges.

On the other hand, *RAB8B* has been shown to function in membrane trafficking in developing hippocampal neurons with loss of *RAB8B* expression shown to impair neurite outgrowth[334]. Additionally, there is evidence of increased expression of *RAB8B* associated with the presence of a specific risk variant in Schizophrenia patients[335]. Although its function in microglia has not been investigated, a study of two proteins in the same family, *Rab20* and *Rab32*, implicates a role for Rab GTPases in generating an inflammatory response in

microglia. Notably, proinflammatory responses of microglia have previously been implicated in the pathogenesis of schizophrenia[336]. Considering that *RAB8B's* expression appears to be temporally regulated during a period of susceptibility for schizophrenia, further investigation of this gene in microglia is warranted.

The DESeq2 and Psupertime analysis conducted here did not focus on the changes in gene expression between any two of the compared epochs. This could be a focus of future work to determine which of the epochs show the greatest cell type-specific differences over the course of brain maturation.

4.4.2. IDEAS analysis

The IDEAS method[222] was used to identify genes that differ in expression level between the broad paediatric and adult groups. In comparison to the DESeq2 analysis, where the samples were grouped into 5 epochs, the statistical power of the IDEAS analysis was greater due to the larger sample size per group allowing for multiple DEGs to be identified even in cell types with very few nuclei. Additionally, IDEAS was designed to identify DEGs that differ between groups in either their expression level or their variance, unlike DESeq2 and Psupertime which were only designed to identify DEGs that change in their expression level.

Notably, in the Astro L1-2 FGFR3 GFAP population, several genes which were upregulated in paediatric samples compared to adult samples were associated with phosphorylation which is a well-established mechanism controlling various processes, including cell cycle progression[337]. Thus, the increased expression of these genes in paediatric versus adult astrocytes may contribute to various developmental processes of astrocytes such as cell proliferation which, in the rat brain, has been found to be limited to a short temporal window in the early postnatal years[338]. In addition to a role in phosphorylation, several genes that showed this pattern of expression, such as *SLC6A1,* were associated with Schizophrenia and temporal lobe epilepsy. Interestingly, a previous study has reported astrocytic dysregulation of *SLC6A1* in the pathogenesis of epilepsy[339], and its differential expression observed here may contribute to age-specific mechanisms of epileptogenesis[340].

In Endo L2-6 NOSTRIN, genes involved in calcium ion transport were upregulated in paediatric samples versus adult samples. Importantly, calcium signalling in brain endothelial cells has recently been shown to play an essential role in directing blood flow[341]. Thus, the increased expression of calcium ion transporters in endothelial cells in paediatric samples may facilitate increased blood supply which is a requirement during this period[342]. Markedly, several genes belonging to this expression pattern were associated with depression and there is evidence supporting a role for reduced cerebral blood flow in the pathogenesis of major depressive disorder (MDD)[343]. That the change in the level of expression of these genes occurs at the transition from adolescence to young adulthood may place young adults at increased risk of developing MDD.

In OPC L1-6 PDGFRA, genes upregulated in the paediatric samples compared to adults were associated with developmental processes occurring during the post-natal period such as axonal transport and synapse organization. Furthermore, these genes were associated with

neurological conditions including migraine and epilepsy which are known to have differential manifestations with age[340,344] and often have specific ages-of-onset[345]. Notably, children are more prone to seizures[346] which has been attributed to hyperexcitability in the immature brain[347]. Thus, genes such as those found in the described pattern, which carry out normal functions during brain maturation, may place paediatric patients at increased risk of developing seizures[347]. An important consideration in my study is that the tissue comes from epilepsy patients. While the tissue did not come from the suspected epileptic focus, it may well be affected by the epileptic environment. Thus, any genes identified that are implicated in seizures should be interpreted with caution and validation of their expression in non-epileptic tissue is warranted, for example by deconvolving paediatric and adult bulk RNA-seq datasets from Werling et al. (2020)[136].

Another interesting pattern to emerge from the IDEAS analysis included a set of genes upregulated in adults in Exc L2 LAMP5 LTK which was enriched for genes associated with cation channel activity and chemical synaptic transmission, including two serotonin receptors and two gamma-aminobutyric acid (GABA) receptors. Notably, several of these genes including *CACNB2*, *DLG1*, *FGF14*, *HTR1E*, *AMPH*, and *HTR4* have previously been reported to increase in expression over the postnatal period across multiple brain regions based on weighted gene co-expression network analysis performed by Kang et al. (2011) on bulk RNA-seq profiles[119]. In Kang et al. (2011), these genes formed part of the same M2 module and, in agreement with our study, were enriched for terms related to calcium signalling and synaptic transmission[119]. Through our single-cell analysis, this expression signature can now be confidently mapped to excitatory neurons. Considering these genes all function in influencing neuronal excitability, the increase in their expression with age suggests a possible maturation in the firing properties of this cell type.

In addition to genes upregulated in Exc L2 LAMP5 LTK with age, there were also genes showing an opposing trajectory, including 5 genes belonging to the same family of solute carrier proteins (SLCs). SLCs have various functions in neurons related to cell type-specific neurotransmitter transport[348] and have been implicated in various neurological conditions including major depressive disorder, epilepsy, and schizophrenia[349–351]. Consequently, they represent important drug targets[352] and their altered expression between paediatric and adult patients seen here could explain differential responses to treatment observed between these groups in conditions such as epilepsy[353]. More than 20 genes in the pattern of age-dependent downregulated genes in Exc L2 LAMP5 LTK were associated with AD, including *BACE1* which has previously been shown to be highly expressed in the post-natal period in mouse motor neurons[354]. There is further evidence that it may play various roles in postnatal neuronal development including in axon guidance[355] and myelination[356] whilst contributing to synaptic plasticity in the adult brain[357]. The identification of AD-related genes that are downregulated with age in Exc L2 LAMP5 LTK agrees with the DESeq2 and Psupertime consensus analysis which also found two AD-related genes, *FNBP1L* and *APOE,* to be downregulated with age in Exc L2 LAMP5 LTK.

MEF2C was also amongst the genes decreasing in its expression trajectory with age in Exc L2 LAMP5 LTK. This agrees with its described expression profile by Li et al. (2018) who profiled the transcriptomes of post-mortem human brain tissue samples spanning embryonic development to adulthood using bulk RNA-seq for most samples and scRNA-seq for a subset of embryonic and adult samples[90]. They used weighted gene correlation network analysis to identify modules of co-expressed genes, including genes showing changes in temporal expression trajectories. *MEF2C* formed part of a module which was enriched in excitatory neurons[90]. Its function in the nervous system is primarily a transcriptional repressor[358] and dysregulation of this gene has been implicated in numerous neuropsychiatric disorders including Autism Spectrum Disorder[359], Schizophrenia[360], Epilepsy[361], and Major Depressive Disorder[362]. In mice, knockout of *Mef2c* in hippocampal excitatory neurons resulted in reduced excitatory synaptic transmission in layer 2/3 pyramidal neurons and increased inhibitory synaptic transmission, with concordant reduction in glutamatergic synaptic density and increases in GABAergic synaptic density[358]. Markedly, in the post-natal brain, *MEF2C* may function in synaptic pruning of excitatory synapses[360,363,364], in contrast to its pre-natal role in synaptogenesis of excitatory neurons. The discrepancy in *MEF2C*'s effect on excitatory neurons between the pre- and post-natal brain could be due to a switch in its functionality over this period or it may have cell-type specific effects on sub-types of excitatory neurons which were not previously assessed[358]. Overall, its expression pattern in our study corroborates the findings from previous studies and shows a trajectory coinciding with the period of synaptic pruning, lending support to its role in this process.

Reassuringly, several genes which decreased with age in Inh L2-4 PVALB WFDC2 were found to be downregulated in the brain according to the GTEx database. This included the GABA B receptor subunit, *GABBR2,* which forms a heterodimer with *GABBR1*[365] and is expressed in both excitatory and inhibitory neurons[366]. It functions pre-synaptically in the secretion of neurotransmitters including GABA and glutamate, and post-synaptically in inhibiting neuronal activity[367,368]. Fukui et al. (2011) examined its mRNA expression level in the mouse brain at three post-natal stages and found it to remain relatively constant from birth to adulthood[369]. However, this was performed on whole tissue sections and so its expression level within specific cell types was masked[369]. Importantly, *GABBR2* has been implicated in various neurodevelopmental disorders including autism spectrum disorder[370], temporal lobe epilepsy[371], and several mood disorders[372], warranting further investigation of its function in Inh L2-4 PVALB WFDC2.

Lastly, in the Inh L3-5 SST ADGRG6 population, a set of genes that was downregulated in adults was also enriched for GTEx terms of genes downregulated in the brain with age, including the zinc finger transcription factor, *MYT1L*. Similarly to *GABBR2, MYTL1* has been associated with numerous neurodevelopmental disorders including intellectual disability, autism spectrum disorder, and schizophrenia[373]. In agreement with its expression pattern in Inh L3-5 SST ADGRG6, it has previously been shown to decrease in expression with age in both the mouse and human brain[374]. An extensive characterization of its function in the

mouse brain through knockdown analysis highlighted its importance in the maturation of neurons, with disruptions in both the structural and physiological properties of neurons observed[374]. Notably however, the loss of function of *MYTL1* had differential effects in the developing mouse brain compared to the adult mouse brain, including binding to different chromatin targets[374]. This may in part be explained by changes in *MYTL1* expression with age since the occupancy of transcription factors at binding sites depends on their concentration[375].

In addition to *MYTL1*, other genes downregulated with age in Inh L3-5 SST ADGRG6 included, *ABCD2, PEX6*, and *HACL1* which were associated with protein targeting to peroxisome. The decrease in expression of these genes with age may result in reduced degradation of damaged proteins with age and consequently place older individuals at risk of cognitive decline[376]. Interestingly, the genes in this pattern were also associated with presenile dementia, implicating Inh L3-5 SST ADGRG6 in the pathogenesis of this condition.

Overall, the sets of differential expression analyses are informative in identifying signatures of gene expression within specific cell types that are implicated in the process of normal brain maturation and may contribute to neurological conditions. If investigated further, the analyses will be useful in identifying putative cell type-specific drug targets as well as appropriate time periods for therapeutic intervention to treat neurological disorders[90,136].

4.4.3. Proportion analysis

In parallel to assessing changes in the level of expression of the consensus genes, changes in the proportion of nuclei expressing the genes were also evaluated. This revealed a general corresponding trend between the level of expression and the proportion of nuclei expressing the gene with age. Considering the high rate of drop out events and the random nature of sampling nuclei[134,210], one cannot be confident that a difference in the proportion of nuclei expressing a gene between groups represents a true difference in the number of nuclei expressing the gene. Nevertheless, assuming that the probability of sampling a gene increases with an increase in its expression level, an increase in the proportion of nuclei expressing a gene between groups may be an indicator of an increase in its level of expression. Thus, this parameter could be useful as an alternative method for examining differences in gene expression between groups.

To test whether this parameter could be useful in identifying relevant genes distinguishing paediatric and adult samples, a hypothesis test was performed to compare the proportion of nuclei expressing each gene per cell type between the two sample groups. Intriguingly, this revealed a significant difference for thousands of genes in the Exc L2 LAMP5 LTK population with almost no genes identified in other cell types, suggesting a potential artefact. Nevertheless, assuming that this is a real effect, it may be due to a lower dropout rate in this cell type compared to most other cell types as indicated by the higher number of genes and transcripts detected for this cell type. Alternatively, for most cell types, the absolute differences in expression levels between paediatric and adult samples may be insufficient to be detected by this measure. The genes which showed a significant change in

the proportion of Exc L2 LAMP5 LTK nuclei expressing them all showed an increase in adults compared to paediatric samples, as well as a general increase in the level of expression, suggesting widespread increases in transcription with age. Further investigation is warranted to determine whether this observation is an artefact.

GSEA implicated a role for the genes that increased in the proportion analysis in synaptic signalling and protein phosphorylation – a mechanism required for synaptic plasticity. Markedly, *HCN1* and *SCN1A* were amongst the significant genes. The *HCN1* protein forms a potassium channel which decreases membrane excitability[377] and its expression in CA1 hippocampal neurons has been shown to reduce synaptic plasticity[378]. Importantly, in excitatory pyramidal neurons, the hyperpolarization-activated current generated by the *HCN1* channel has been shown to significantly increase in its amplitude over the post-natal period in the mouse prefrontal cortex[379]. This seems to agree with the observed increased expression of *HCN1* in Exc L2 LAMP5 LTK as the brain matures. In contrast, in inhibitory neurons the amplitude of the hyperpolarization-activated current decreases during post-natal brain maturation[379], highlighting the importance of examining the expression and function of ion channels within specific cell types. On the other hand, *SCN1A* is a voltage-gated sodium channel whose expression should increase neuronal excitability[275]. Its expression has been found to increase during the post-natal period in whole cortical tissue samples from rodents and humans[380,381]. Additionally, loss of function of *SCN1A* in GABAergic neurons is the most common case of Dravet Syndrome[382] - a severe epileptic encephalopathy. Interestingly, gain of function mutations in *HCN1* have also been implicated in Dravet Syndrome[383], presumably by also affecting inhibitory neurons. How the increase in expression of *HCN1* and *SCN1A* in adult Exc L2 LAMP5 LTK neurons alters the firing properties and plasticity of these circuits warrants further investigation, especially considering their seemingly contradictory functions.

4.5. Long non-coding RNA analysis

A growing body of evidence suggests that lncRNAs represent an important class of molecules as regulators of cell type-specific functions in the brain[154,156,158,219]. However, the functions of most lncRNAs remain unknown. This prompted the *in silico* investigation into the functions of two lncRNAs identified in this study.

LINC00499, was selected for further characterization as it showed simultaneous cell-type-specificity for Astro L1-6 FGFR3 SLC14A1 whilst also being upregulated in this cell type with age according to DESeq2 and Psupertime. Investigating possible functions of *LINC00499*, through guilt-by-association (GBA) analysis suggests that it may function in pathways which switch off neurodevelopmental and synaptic functions, whilst switching on immune-related functions as the brain matures. Since astrocytes, oligodendrocytes, and neurons come from the same progenitors[384], it is possible that immature astrocytes have expression signatures reminiscent of neurons and oligodendrocytes. For example, based on GBA for *LINC00499*, *TENM4* showed a directly opposing expression pattern to *LINC00499* in Astro L1-6 FGFR3 SLC14A1 – with high levels of expression in the first epoch which were subsequently downregulated. *TENM4* expression has previously been associated with oligodendrocyte

differentiation[385], axon guidance[386], synapse adhesion[387], and increased astrocyte territory occupation[388] which is a phenotype characteristic of the early post-natal brain[389]. Other synaptic genes such as *GRID2*[390] and *NLGN1*[391] were also negatively correlated with *LINC00499* based on GBA and appeared to be downregulated in Astro L1-6 FGFR3 SLC14A1 with age. Likewise, three genes belonging to the Regulator of G protein signaling family (*RGS6*, *RGS7*, and *RGS20*) also showed an opposing pattern to *LINC00499*, with *RGS7* and *RGS20* being amongst the FasimTarget analysis hits. These genes negatively regulate the G protein-coupled receptor signalling pathway[392], which plays a critical role in mediating astrocyte-neuronal communication, in turn effecting synaptic transmission and plasticity[393]. In contrast to their expression profile, *AQP4*, which is a marker of mature astrocytes[394], was upregulated with age in Astro L1-6 FGFR3 SLC14A which correlates with the expression profile of *LINC00499*.

In support of the interpretation of *LINC00499* serving as a regulatory switch, an exploration of putative DNA binding motifs of *LINC00499* revealed that most hits showed a general opposing expression profile to that of *LINC00499*, suggesting it may act as a negative regulator of gene expression. The top 50 genes (based on the percentage of nuclei expressing the genes) were associated with processes such as neuron development and Wnt signaling which is involved in oligodendrocyte differentiation[395]. This agrees with the hypothesis that increased expression of *LINC00499* with age may switch off neuron and oligodendrocyte-related pathways that are active in the early post-natal period. Importantly, variants at the *LINC00499* locus have been implicated in various neurological conditions including Autism Spectrum Disorder[396], Major Depressive Disorder[397] and differential responses amongst Schizophrenia patients to the drug, paliperidone[398] . Thus, alterations in the expression of *LINC00499* during critical post-natal periods may disrupt normal astrocyte functioning and thereby increase risk for neuropsychiatric conditions.

AC004852.2 is also an lncRNA of interest as it was identified as a novel cell type-specific marker for 10 of the samples in the OPC L1-6 PDGFRA population. GBA analysis indicated a possible role for this gene in synaptic signalling and it was co-expressed with genes such *SCN1A* which, as discussed above, is a voltage gated sodium ion channel that has been implicated in the pathogenesis of various epilepsies (both loss and gain of function mutations)[399,400]. *SCN1A* expression in OPCs has previously been described but its function in these cells remains unknown[401]. In addition to *SCN1A, NOVA1* was also co-expressed with *AC004852.2* and was discovered as a DNA interaction partner according to FasimTarget analysis. *NOVA1* functions in alternative splicing in neurons and was previously characterised as a neuron-specific protein[402]. However, more recent analysis and evidence from this study suggests it is also expressed in OPCs[108,403].

Long non-coding RNAs are increasingly implicated in polygenic neurological disorders including conditions which typically manifest during critical periods post-natally. For example, in a study of major depression using post-mortem human brain specimens, lncRNAs represented almost a third of the DEGs identified between cases and controls[404]. Thus, characterising their expression dynamics in the brain with age and determining the

specific cell types which they are expressed in may help to elucidate the pathogenesis of complex disorders.

4.6. Validation of snRNA-seq analysis results using Visium spatial transcriptomics

The 10X Genomics Visium platform was used to validate the cell type-specific expression of several of the NS-Forest minimal marker genes using a subset of samples and cell types. The Visium datasets were annotated using the merged paediatric and adult snRNA-seq datasets generated in this study in order to render the two sets of transcriptomic data comparable. Where a label transfer approach was used to annotate the snRNA-seq datasets, an independent machine learning method from the cell2location package was used to annotate the Visium datasets. Assessing the expression of four of the NS-Forest minimal marker genes confirmed their expression in the expected cell types. This included *LINC00499* in Astro L1-6 FGFR3 SLC14A1, *AC004852.2* in OPC L1-6 PDGFRA, *APBB1IP* in Micro L1-3 TYROBP, and *SEMA3E* in Exc L5-6 FEZF2 ABO.

LINC00499 expression in the Visium datasets appeared to largely be restricted to the Astro L1-6 FGFR3 SLC14A1 population in agreement with its expression in the snRNA-seq data. Furthermore, the 4-year-old had the fewest number of positive spots followed by the two 15-year-old samples while the 31-year-old had the greatest number of positive spots – in line with the proportion of nuclei expressing *LINC00499* in each of the snRNA-seq samples. Interestingly, most of the spots estimated to have greatest abundance of the Astro L1-6 FGFR3 SLC14A1 cell subtype did not seem to express *LINC00499* indicating that it was only expressed in a subset of Astro L1-6 FGFR3 SLC14A1 cells. This finding could be further validated by analysing *LINC00499* expression alongside known cell type markers in tissue sections using more targeted approaches such as *in situ* hybridization chain reaction[405].

Notably, in the snRNA-seq datasets, *AC004852.2* was expressed at high levels and in a large proportion of nuclei in OPCs across all samples. However, the Visium analysis revealed very few spots expressing these genes for three of the samples, contrary to expectation. This could be due to the lower sensitivity of the Visium platform compared to the snRNA-seq method since Visium is unable to detect single cells but instead captures up to 30 cells per spot[105]. Alternatively, there may be fewer OPC cells in the tissue sections used for Visium compared to the pieces of tissue used to generate snRNA-seq libraries, or the cell2location analysis may have incorrectly estimated the abundance of OPCs in the Visium data.

A previously described marker of microglia, *APBB1IP*[270], was also found to be a marker gene for microglia in the snRNA-seq data where it was largely specific to Micro L1-3 TYROBP and was expressed in a large proportion of nuclei in this population across all snRNA-seq samples. This was validated in the Visium datasets in which it was most highly expressed in Micro L1-3 TYROBP cells compared to cells of other populations across all four samples. Furthermore, it appeared to be expressed in a large proportion of this population as indicated by its colocalization with spots showing greatest estimated abundance of Micro L1-3 TYROBP cells. In both the Visium and snRNA-seq data, *APBB1IP* appeared to also be

expressed in Exc L4-6 FEZF2 IL26, however this was at lower levels than those observed in Micro L1-3 TYROBP. As aforementioned, several other microglial markers in the snRNA-seq data were also expressed at low levels in Exc L4-6 FEZF2 IL26 which seems to suggest that the Exc L4-6 FEZF2 IL26 nuclei were incorrectly annotated. However, the independent validation of *APBB1IP* expression in Exc L4-6 FEZF2 IL26 in the Visium analysis, which was annotated using a different approach to the label transfer method used in the snRNA-seq analysis, favours an alternative interpretation that this excitatory neuron population does indeed express microglial genes, albeit at lower levels than in microglia.

SEMA3E was identified as a marker gene for Exc L5-6 FEZF2 ABO in 10 of the snRNA-seq datasets and showed high cell-type-specificity for this cell type. Markedly, Hodge et al. (2019)[109] identified *SEMA3E* as a minimal marker of Exc L4-6 RORB *SEMA3E* not Exc L5-6 FEZF2 ABO. However, in the Visium datasets, it was expressed at high levels in Exc L5-6 FEZF2 ABO whilst being expressed at much lower levels in Exc L4-6 RORB SEMA3E, validating it as a cell type-specific marker for Exc L5-6 FEZF2 ABO in our datasets. That *SEMA3E* appears to be a novel marker of Exc L5-6 FEZF2 ABO in our study and not in Hodge et al. (2019) may be an indication that the current draft human brain atlas does not fully capture human brain transcriptomic diversity. Thus, to achieve a more comprehensive human brain cell atlas, the integration of the current atlas with a greater number of diverse datasets, including samples from diverse populations, age groups, and sexes is warranted. Moreover, the inclusion of a greater number of datasets that have been processed in different laboratories will help to distinguish technical effects, such as sample storage methods, sample processing, and data analysis methods, from biological effects.

4.7. Limitations and future directions

While this research has provided insight into the cell type-specific gene expression changes that may be occurring during brain maturation, a clear limitation of this study is the small sample size which limits the generalizability of the findings to the larger population. Importantly, a large source of variation between samples was the variable distinguishing the publicly available datasets from our datasets. This effect may be due to the single cell chemistry platform used as the sequencing depth improved in the v3.1 (used to generate our datasets) versus v2.0 chemistry (used to generate the publicly available datasets)[134]. However, contrary to expectation, a greater average sequencing depth was observed in the publicly available datasets compared to our datasets which may be due to fewer barcodes being sequenced in the publicly available datasets compared to our datasets. This would allow for greater depth per barcode in the publicly available datasets which may explain the batch effect observed. To try minimize the impact of this batch variable (referred to as single cell chemistry platform in the metadata), its effect was modelled in the DESeq2 and IDEAS pipelines and adjusted accordingly. Future studies can randomly down-sample the number of reads to that of the sample with the fewest reads to account for differences in sequencing depth between samples and minimise batch effects[406]. However, this poses the risk of losing valuable information.

The sex of the donors was also a potential confounding variable especially since there was a bias between groups with only one male in the adult age group compared to 3 males in the paediatric group. This was not regressed out with the single cell chemistry variable as the inclusion of too many covariates in the regression model could inadvertently regress out biological variation due to the independent variable of interest[209]. Nonetheless, each of the resulting gene expression patterns were visually inspected for sex-specific effects. Overall, it appeared that sex was not a major confounding variable but future studies such as this can specifically model its effect to identify temporally regulated genes that are specific to a particular sex. This may provide insight into sex-specific differences in neurological conditions, for example why depression and anxiety disorders are more common in females than males[407,408]. Overall, increasing the sample size will minimize the effect of confounding variables whilst increasing the power to identify true differences between the age groups. There are currently two research groups funded by the Chan Zuckerberg foundation which are working towards generating paediatric brain cell atlases[409]. In the coming years, these can be integrated with our datasets to further explore the generalisability of our findings.

Another critical consideration is the use of tissue obtained from epilepsy patients. As mentioned above, the tissue did not come from the suspected epileptic focus. However, considering that many epilepsies have a large genetic component[361], it is possible that there are underlying epileptogenic signatures affecting the tissue. It is also possible that some of the transcriptomic differences observed between the paediatric and adult datasets represent age-dependent differences in the pathogenesis of epilepsy and not differences which would ordinarily be observed between healthy subjects. It is reassuring that the current differential expression analyses revealed numerous genes which have previously been shown to change with age in bulk RNA-seq datasets generated using non-epileptic tissue. This suggests that we have successfully detected temporal signatures resembling that of normal tissue. Moreover, the use of tissue from epilepsy patients in this study is a valuable trade-off for the opportunity of using ante-mortem brain tissue which is less susceptible to RNA degradation and artefacts relating to the post-mortem period[133,410,411]. Nonetheless, a recommendation for future studies would be to generate snRNA-seq datasets using both the epilepsy focal tissue and the surrounding tissue from the same patients to determine the extent to which these sources of tissue are the same. Alternatively, the study could be repeated using post-mortem paediatric and adult tissue to see if the results in this study can be replicated.

The use of nuclei in this study instead of cells was necessary as the tissue had to be frozen for logistical reasons and use of cells from frozen tissue is not a viable option[128]. Notably, the nuclear fraction provides a limited picture of total transcript diversity since it only represents a portion of transcripts in the entire cellular fraction, with the expression of some genes such as microglial activation genes appearing to differ between these two fractions[129,255]. Nonetheless, there is evidence that snRNA-seq and scRNA-seq are comparable in their ability to resolve cell types, even transcriptionally similar neuronal cell types[129], suggesting that the use of nuclei is not a major limitation.

In terms of the interpretation of the differential expression analysis, GSEA was performed using Enrichr to identify putative functions of the genes. This method was chosen due to the variety of databases available including the GTEx Aging Signatures database[299], which was relevant to the context of this study. However, there are various other approaches to carrying out an enrichment analysis which could be applied to these datasets in a future study[228]. For example the Broad Institute's GSEA method[229] could be applied which takes a ranked list of genes as input and has been suggested to outperform hypergeometric methods such as Enrichr[285]. Markedly, many of the terms identified in this study did not reach significance after adjusting for multiple testing. Nonetheless, both the significant and non-significant terms provide preliminary insight into putative functions of the genes which can be used to generate hypotheses that can be tested experimentally within specific cell types of interest.

Another important consideration of the differential expression analyses is that many of the significant DEGs were expressed in a low proportion of nuclei in the cell populations under investigation (i.e. < 20% of nuclei). As we did not saturate sequencing depth, it is unclear whether these genes are indeed only expressed in a fraction of nuclei in the population or if they were only detected in a fraction of the nuclei due to some nuclei not being sequenced to a great enough depth. While the data was normalised for sequencing depth, it is possible that differences in sequencing depth were not fully corrected for and that some DEGs may be artefacts due to the random nature of sampling. This may be improved by increasing the sequencing depth and sample size. Additionally, future studies can use a more stringent threshold to exclude genes that are expressed in a small percentage of nuclei. As this was an exploratory study, I chose not to use stringent exclusion criteria to avoid the possibility of losing genes of interest.

With regards to the lncRNA analysis, the characterization performed here was very coarse as it was only performed to generate hypotheses as to the putative functions of the selected molecules. However, to refine the characterization performed here and provide robust evidence of their functions, further characterization is warranted such as identifying global DNA interaction partners of the lncRNAs using GRID-seq[412], identifying global protein interaction partners using methods such as RNA affinity purification followed by mass spectrometry[413], and performing knockout or overexpression analyses in specific cell lines[169]. Notably, the 3' end sequencing method used by 10x Genomics is designed to target polyadenylated transcripts and thus some lncRNAs may have gone undetected in this study since not all lncRNAs are poly-adenylated.

Lastly, the analysis performed here provides a picture of the molecular changes driving post-natal human brain maturation but there are many levels of genetic and epigenetic regulation not captured by only examining the transcriptome. For example, epigenetic modifications such as methylation can alter which genes are expressed which requires alternative techniques to detect such as bisulfite sequencing in single cells[414]. Additionally, translation does not necessarily recapitulate transcription and so validation of the expression of protein-coding genes using methods such as immunohistochemistry is

warranted. Another consideration is that post-transcriptional splicing can greatly influence the levels of specific isoforms of genes[415], adding another layer of complexity to gene regulatory networks since different isoforms can have different regulatory roles[416]. The 10X Genomics method used here only profiles the 3' end of transcripts and so isoform information is lost. However, alternative snRNA-seq techniques, such as Smart-seq3, capture full length transcripts allowing one to study cell type-specific differences in the levels of transcript isoforms[417]. This protocol is in the pipeline in our laboratory which will allow us to examine whether certain gene isoforms are differentially expressed with age. Alongside this, another project in our laboratory has generated single-nucleus Assay for Transposase Accessible Chromatin (snATAC) sequencing profiles using the same samples as the snRNA-seq datasets. By integrating this epigenetic analysis with the snRNA-seq analysis we may gain insight into how chromatin regulation influences age-dependent gene expression dynamics.

4.8. Conclusion

The study expands on the literature by testing the generalisability of current draft MTG cell atlas to novel samples. While the draft atlas could successfully be used as a reference to annotate our datasets, distinct differences between our datasets and the reference dataset were apparent including differences in the number of cell types identified and differences in marker genes distinguishing cell types. This observation warrants the inclusion of a greater number of diverse datasets into the reference human brain cell atlas in order to have a comprehensive reference atlas which can be applied to samples of diverse demographics. Markedly, the datasets generated in this study serve as a locally and globally relevant resource to explore differences between the paediatric and adult human brain at single cell resolution. Through the analysis carried out here, I was able to identify transcriptomic differences between the age groups, including identifying putative marker genes of specific cell types capable of distinguishing paediatric from adult brains. Additionally, genes changing in their level of expression with age or in the proportion of nuclei expressing them were identified for various neuronal and non-neuronal cell populations. Notably, the differential expression analysis revealed that numerous age-dependent differences in gene expression were cell type and subtype-specific, affirming the importance of exploring gene expression dynamics during postnatal human brain maturation at single cell resolution. The study reinforces the relevance of non-coding genes in the brain and highlights their possible contribution to brain cell-type diversity. Importantly, this research may contribute to our understanding of neurological disorders, especially conditions which manifest differently between children and adults.

References

1. de Graaf-Peters, V. B. & Hadders-Algra, M. Ontogeny of the human central nervous system: What is happening when? *Early Hum Dev* **82**, 257–266 (2006).

2. Stiles, J. & Jernigan, T. L. The basics of brain development. *Neuropsychology Review* Preprint at https://doi.org/10.1007/s11065-010-9148-4 (2010).

3. Azevedo, F. A. C. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol* **513**, 532–541 (2009).

4. Doyle, D. The Fine Structure of the Nervous System: The Neurons and Supporting Cells. *J Neurol Neurosurg Psychiatry* **41**, (1978).

5. Bakken, T. *et al.* Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* **18**, 7–16 (2017).

6. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).

7. Cadwell, C. R., Sandberg, R., Jiang, X. & Tolias, A. S. Q&A: Using Patch-seq to profile single cells. *BMC Biol* **15**, 1–7 (2017).

8. Khasawneh, A., Garling, R. & Harris, C. Cerebrospinal fluid circulation: What do we know and how do we know it? *Brain Circ* **4**, 14 (2018).

9. Jessell, T. M. & Sanes, J. R. Development: The decade of the developing brain. *Curr Opin Neurobiol* **10**, 599–611 (2000).

10. Bystron, I., Rakic, P., Molnár, Z. & Blakemore, C. The first neurons of the human cerebral cortex. *Nat Neurosci* **9**, 880–886 (2006).

11. Bystron, I., Blakemore, C. & Rakic, P. Development of the human cerebral cortex: Boulder Committee revisited. *Nat Rev Neurosci* **9**, 110–122 (2008).

12. Tau, G. Z. & Peterson, B. S. Normal Development of Brain Circuits. *Neuropsychopharmacology* **35**, 147–168 (2010).

13. Purves, D., Augustine, G. & Fitzpatrick, D. Excitatory and Inhibitory Postsynaptic Potentials. in *Neuroscience* (ed. Editors) (Sinauer Associates, 2001).

14. Marin, O., Valiente, M., Ge, X. & Tsai, L.-H. Guiding Neuronal Cell Migrations. *Cold Spring Harb Perspect Biol* **2**, a001834–a001834 (2010).

15. Cooper, J. A. A mechanism for inside-out lamination in the neocortex. *Trends Neurosci* **31**, 113–119 (2008).

16. Sakai, N. & Kaprielian, Z. Guidance of longitudinally projecting axons in the developing central nervous system. *Front Mol Neurosci* **5**, (2012).

17. Jan, Y.-N. & Jan, L. Y. Branching out: mechanisms of dendritic arborization. *Nat Rev Neurosci* **11**, 316–328 (2010).

18. Singh, R. *et al.* Fibroblast Growth Factor 22 Contributes to the Development of Retinal Nerve Terminals in the Dorsal Lateral Geniculate Nucleus. *Front Mol Neurosci* **4**, (2012).

19. Mrzljak, L., Uylings, H. B. M., Van Eden, G. G. & Judáš, M. Chapter 9 Neuronal development in human prefrontal cortex in prenatal and postnatal stages. in 185–222 (1991). doi:10.1016/S0079-6123(08)62681-3.

20. Koenderink, M. J. Th., Uylings, H. B. M. & Mrzljak, L. Postnatal maturation of the layer III pyramidal neurons in the human prefrontal cortex: a quantitative Golgi analysis. *Brain Res* **653**, 173–182 (1994).

21. Michel, A. E. & Garey, L. J. The development of dendritic spines in the human visual cortex. *Hum Neurobiol* **3**, 223–7 (1984).

22. Koenderink, M. J. Th. & Uylings, H. B. M. Postnatal maturation of layer V pyramidal neurons in the human prefrontal cortex. A quantitative Golgi analysis. *Brain Res* **678**, 233–243 (1995).

23. Becker, L. E., Armstrong, D. L., Chan, F. & Wood, M. M. Dendritic development in human occipital cortical neurons. *Developmental Brain Research* **13**, 117–124 (1984).

24. Tallinen, T. *et al.* On the growth and form of cortical convolutions. *Nat Phys* **12**, 588–593 (2016).

25. Molofsky, A. V. & Deneen, B. Astrocyte development: A Guide for the Perplexed. *Glia* **63**, 1320–1329 (2015).

26. Kimelberg, H. K. Functions of Mature Mammalian Astrocytes: A Current View. *The Neuroscientist* **16**, 79–106 (2010).

27. Nave, K.-A. & Werner, H. B. Myelination of the Nervous System: Mechanisms and Functions. *Annu Rev Cell Dev Biol* **30**, 503–533 (2014).

28. Williamson, J. M. & Lyons, D. A. Myelin Dynamics Throughout Life: An Ever-Changing Landscape? *Front Cell Neurosci* **12**, (2018).

29. Yakovlev, P. & Lecours, A. The myelogenetic cycles of regional maturation of the brain. In: Resional development of the brain in early life. *Oxford: Blackwell* 3–70 (1967).

30. Young, K. M. *et al.* Oligodendrocyte Dynamics in the Healthy Adult CNS: Evidence for Myelin Remodeling. *Neuron* **77**, 873–885 (2013).

31. Scantlebury, N. *et al.* Relations between White Matter Maturation and Reaction Time in Childhood. *Journal of the International Neuropsychological Society* **20**, 99–112 (2014).

32. Lin, S. & Bergles, D. E. Synaptic signaling between GABAergic interneurons and oligodendrocyte precursor cells in the hippocampus. *Nat Neurosci* **7**, 24–32 (2004).

33. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat Neurosci* **23**, 771–781 (2020).

34. Nishiyama, A., Yu, M., Drazba, J. A. & Tuohy, V. K. Normal and reactive NG2+ glial cells are distinct from resting and activated microglia. *J Neurosci Res* **48**, 299–312 (1997).

35. Hamilton, N., Vayro, S., Wigley, R. & Butt, A. M. Axons and astrocytes release ATP and glutamate to evoke calcium signals in NG2-glia. *Glia* **58**, 66–79 (2010).

36. Ziskin, J. L., Nishiyama, A., Rubio, M., Fukaya, M. & Bergles, D. E. Vesicular release of glutamate from unmyelinated axons in white matter. *Nat Neurosci* **10**, 321–330 (2007).

37. Tomassy, G. S. *et al.* Distinct Profiles of Myelin Distribution Along Single Axons of Pyramidal Neurons in the Neocortex. *Science (1979)* **344**, 319–324 (2014).

38. Ginhoux, F. *et al.* Fate Mapping Analysis Reveals That Adult Microglia Derive from Primitive Macrophages. *Science (1979)* **330**, 841–845 (2010).

39. Nimmerjahn, A., Kirchhoff, F. & Helmchen, F. Resting Microglial Cells Are Highly Dynamic Surveillants of Brain Parenchyma in Vivo. *Science (1979)* **308**, 1314–1318 (2005).

40. Benarroch, E. E. Microglia: Multiple roles in surveillance, circuit shaping, and response to injury. *Neurology* **81**, 1079–1088 (2013).

41. Hammond, T. R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* **50**, 253-271.e6 (2019).

42. Paolicelli, R. C. *et al.* Synaptic pruning by microglia is necessary for normal brain development. *Science (1979)* (2011) doi:10.1126/science.1202529.

43. Weinhard, L. *et al.* Microglia remodel synapses by presynaptic trogocytosis and spine head filopodia induction. *Nat Commun* (2018) doi:10.1038/s41467-018-03566-5.

44. Afroz, S., Parato, J., Shen, H. & Smith, S. S. Synaptic pruning in the female hippocampus is triggered at puberty by extrasynaptic GABAA receptors on dendritic spines. *Elife* **5**, (2016).

45. Birnbaum, R. & Weinberger, D. R. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat Rev Neurosci* **18**, 727–740 (2017).

46. Shaw, P., Gogtay, N. & Rapoport, J. Childhood psychiatric disorders as anomalies in neurodevelopmental trajectories. *Hum Brain Mapp* **31**, 917–925 (2010).

47. Cowan, W. M., Fawcett, J. W., O'Leary, D. D. M. & Stanfield, B. B. Regressive Events in Neurogenesis. *Science (1979)* **225**, 1258–1265 (1984).

48. Blaschke, A. J., Staley, K. & Chun, J. Widespread programmed cell death in proliferative and postmitotic regions of the fetal cerebral cortex. *Development* **122**, 1165–74 (1996).

49. Thomaidou, D., Mione, M. C., Cavanagh, J. F. R. & Parnavelas, J. G. Apoptosis and Its Relation to the Cell Cycle in the Developing Cerebral Cortex. *The Journal of Neuroscience* **17**, 1075–1085 (1997).

50. Goldman-Rakic, P. S. Development of cortical circuitry and cognitive function. *Child Dev* **58**, 601–22 (1987).

51. Peter R., H. Synaptic density in human frontal cortex — Developmental changes and effects of aging. *Brain Res* **163**, 195–205 (1979).

52. Bourgeois, J. & Rakic, P. Changes of synaptic density in the primary visual cortex of the macaque monkey from fetal to adult stage. *The Journal of Neuroscience* **13**, 2801–2820 (1993).

53. Navlakha, S., Barth, A. L. & Bar-Joseph, Z. Decreasing-Rate Pruning Optimizes the Construction of Efficient and Robust Distributed Networks. *PLoS Comput Biol* (2015) doi:10.1371/journal.pcbi.1004347.

54. Hebb, D. O. *The Organization of Behaviour: A Neuropsychological Theory*. (Wiley, 1949).

55. Bliss, T. V. P. & Gardner-Medwin, A. R. Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path. *J Physiol* **232**, 357–374 (1973).

56. Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human cerebral cortex. *Journal of Comparative Neurology* (1997) doi:10.1002/(SICI)1096-9861(19971020)387:2<167::AID-CNE1>3.0.CO;2-Z.

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

58. Fu, M. & Zuo, Y. Experience-dependent structural plasticity in the cortex. *Trends Neurosci* **34**, 177–187 (2011).

59. Hartshorne, J. K., Tenenbaum, J. B. & Pinker, S. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* **177**, 263–277 (2018).

60. Odgers, C. L. *et al.* Is It Important to Prevent Early Exposure to Drugs and Alcohol Among Adolescents? *Psychol Sci* **19**, 1037–1044 (2008).

61. Schmidt, S. *et al.* Experience-dependent structural plasticity in the adult brain: How the learning brain grows. *Neuroimage* **225**, 117502 (2021).

62. Horton, J. C. & Hocking, D. R. An adult-like pattern of ocular dominance columns in striate cortex of newborn monkeys prior to visual experience. *Journal of Neuroscience* (1996) doi:10.1523/jneurosci.16-05-01791.1996.

63. Wong, R. O. L., Meister, M. & Shatz, C. J. Transient period of correlated bursting activity during development of the mammalian retina. *Neuron* (1993) doi:10.1016/0896-6273(93)90122-8.

64. Antón-Bolaños, N. *et al.* Prenatal activity from thalamic neurons governs the emergence of functional cortical maps in mice. *Science (1979)* (2019) doi:10.1126/science.aav7617.

65. Goodman, C. S. & Shatz, C. J. Developmental mechanisms that generate precise patterns of neuronal connectivity. *Cell* Preprint at https://doi.org/10.1016/S0092-8674(05)80030-3 (1993).

66. Pan, Y. & Monje, M. Activity shapes neural circuit form and function: A historical perspective. *Journal of Neuroscience* Preprint at https://doi.org/10.1523/JNEUROSCI.0740-19.2019 (2020).

67. Le Vay, S., Wiesel, T. N. & Hubel, D. H. The development of ocular dominance columns in normal and visually deprived monkeys. *Journal of Comparative Neurology* (1980) doi:10.1002/cne.901910102.

68. Jones, KennethL. & Smith, DavidW. RECOGNITION OF THE FETAL ALCOHOL SYNDROME IN EARLY INFANCY. *The Lancet* **302**, 999–1001 (1973).

69. Cizeron, M. *et al.* A lifespan program of mouse synaptome architecture. *bioRxiv* Preprint at https://doi.org/10.1101/838458 (2019).

70. Huttenlocher, P. R. Morphometric study of human cerebral cortex development. *Neuropsychologia* (1990) doi:10.1016/0028-3932(90)90031-I.

71. Vanhatalo, S. & Kaila, K. Development of neonatal EEG activity: From phenomenology to physiology. *Semin Fetal Neonatal Med* **11**, 471–478 (2006).

72. Marzvanyan, A. & Alhawaj, A. F. *Physiology, Sensory Receptors*. (StatPearls Publishing, 2021).

73. Robinson, K. Implications of developmental plasticity for the language acquisition of deaf children with cochlear implants. *Int J Pediatr Otorhinolaryngol* **46**, 71–80 (1998).

74. Hubel, D. H., Wiesel, T. N. & LeVay, S. Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **278**, 377–409 (1977).

75. Payne, B. R. & Lomber, S. G. Review: Plasticity of the Visual Cortex after Injury: What's Different about the Young Brain? *The Neuroscientist* **8**, 174–185 (2002).

76. Jones, T. A. & Greenough, W. T. Ultrastructural Evidence for Increased Contact between Astrocytes and Synapses in Rats Reared in a Complex Environment. *Neurobiol Learn Mem* **65**, 48–56 (1996).

77. MARKHAM, J. A. & GREENOUGH, W. T. Experience-driven brain plasticity: beyond the synapse. *Neuron Glia Biol* **1**, 351–363 (2004).

78. Black, J. E., Sirevaag, A. M. & Greenough, W. T. Complex experience promotes capillary formation in young rat visual cortex. *Neurosci Lett* **83**, 351–355 (1987).

79. Stiles, J. *The fundamentals of brain development: Integrating nature and nurture.* (Harvard University Press, 2008).

80. Chugani, H. T., Phelps, M. E. & Mazziotta, J. C. Positron emission tomography study of human brain functional development. *Ann Neurol* **22**, 487–497 (1987).

81. Matsuzawa, J. Age-related Volumetric Changes of Brain Gray and White Matter in Healthy Infants and Children. *Cerebral Cortex* **11**, 335–342 (2001).

82. Lebel, C., Walker, L., Leemans, A., Phillips, L. & Beaulieu, C. Microstructural maturation of the human brain from childhood to adulthood. *Neuroimage* **40**, 1044–1055 (2008).

83. Cascio, C. J., Gerig, G. & Piven, J. Diffusion Tensor Imaging. *J Am Acad Child Adolesc Psychiatry* **46**, 213–223 (2007).

84. Gogtay, N. *et al.* Dynamic mapping of human cortical development during childhood through early adulthood. *Proc Natl Acad Sci U S A* (2004) doi:10.1073/pnas.0402680101.

85. Gennatas, E. D. *et al.* Age-Related Effects and Sex Differences in Gray Matter Density, Volume, Mass, and Cortical Thickness from Childhood to Young Adulthood. *The Journal of Neuroscience* **37**, 5065–5073 (2017).

86. Saunders, N. R., Liddelow, S. A. & Dziegielewska, K. M. Barrier Mechanisms in the Developing Brain. *Front Pharmacol* **3**, (2012).

87. Silbereis, J. C., Pochareddy, S., Zhu, Y., Li, M. & Sestan, N. The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron* **89**, 248–268 (2016).

88. Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nat Rev Neurosci* **9**, 947 (2008).

89. Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**, 1117–1125 (2018).

90. Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science (1979)* **362**, (2018).

91. Keshavan, M. S., Giedd, J., Lau, J. Y. F., Lewis, D. A. & Paus, T. Changes in the adolescent brain and the pathophysiology of psychotic disorders. *Lancet Psychiatry* **1**, 549–558 (2014).

92. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* (2007).

93. Solmi, M. *et al.* Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry 2021 27:1* **27**, 281–295 (2021).

94. 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).

95. Ackermann, S. & van Toorn, R. Managing first-time seizures and epilepsy in children. *Continuing Medical Education* **30**, (2012).

96. Naidoo, D. Traumatic brain injury: The South African landscape. *South African Medical Journal* **103**, 613 (2013).

97. Kaku, M. *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind*. (Doubleday, 2014).

98. Ackerman, S. *Discovering the Brain*. (National Academies Press (US), 1992).

99. Callaway, E. M. *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, (2021).

100. Haniffa, M. *et al.* A roadmap for the Human Developmental Cell Atlas. *Nature* vol. 597 Preprint at https://doi.org/10.1038/s41586-021-03620-1 (2021).

101. Regev, A. *et al.* Science Forum: The Human Cell Atlas. *Elife* (2017).

102. Joglekar, A. *et al.* A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463 (2021).

103. Zhang, M. *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).

104. Booeshaghi, A. S. *et al.* Isoform cell-type specificity in the mouse primary motor cortex. *Nature* **598**, 195–199 (2021).

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

106. Di Bella, D. J. *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature* **595**, 554–559 (2021).

107. Langseth, C. M. *et al.* Comprehensive in situ mapping of human cortical transcriptomic cell types. *Commun Biol* **4**, (2021).

108. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).

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

110. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* (2018) doi:10.1038/nbt.4038.

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

112. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

113. Velmeshev, D. *et al.* Single-cell genomics identifies cell type–specific molecular changes in autism. *Science (1979)* (2019) doi:10.1126/science.aav8130.

114. Wang, P., Zhao, D., Lachman, H. M. & Zheng, D. Enriched expression of genes associated with autism spectrum disorders in human inhibitory neurons. *Transl Psychiatry* **8**, 13 (2018).

115. Lau, S.-F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proceedings of the National Academy of Sciences* **117**, 25800–25809 (2020).

116. Johansen, N. *et al.* Inter-individual variation in human cortical cell type abundance and expression. *bioRxiv* 2022.10.07.511366 (2022) doi:10.1101/2022.10.07.511366.

117. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat Neurosci* **24**, 584–594 (2021).

118. Bhaduri, A. *et al.* An atlas of cortical arealization identifies dynamic molecular signatures. *Nature* **598**, 200–204 (2021).

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

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

121. Dönertaş, 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* (2017) doi:10.1038/s41598-017-05927-4.

122. Arneson, D. *et al.* Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat Commun* **9**, 3894 (2018).

123. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37–53 (2016).

124. Dong, X. *et al.* Comprehensive Identification of Long Non-coding RNAs in Purified Cell Types from the Brain Reveals Functional LncRNA in OPC Fate Determination. *PLoS Genet* (2015) doi:10.1371/journal.pgen.1005669.

125. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**, (2014).

126. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* vol. 50 Preprint at https://doi.org/10.1038/s12276-018-0071-8 (2018).

127. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, (2019).

128. Jiang, A. *et al.* Isolated nuclei from frozen tissue are the superior source for single cell RNA-seq compared with whole cells. *bioRxiv* (2023) doi:https://doi.org/10.1101/2023.02.19.529150.

129. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, (2018).

130. Lee, K. *et al.* Human in vitro systems for examining synaptic function and plasticity in the brain. *J Neurophysiol* **123**, 945–965 (2020).

131. Kozareva, V. *et al.* A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219 (2021).

132. Yao, Z. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103–110 (2021).

133. Nagy, C. *et al.* Effects of Postmortem Interval on Biomolecule Integrity in the Brain. *J Neuropathol Exp Neurol* **74**, 459–469 (2015).

134. 10x Genomics. Chromium Single Cell 3' Reagent Kits User Guide (v3.1 Chemistry). https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v31-chemistry (2019).

135. Gawel, D. R. *et al.* A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. *Genome Med* **11**, 47 (2019).

136. 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 (2020).

137. Chicurel, M., Terrian, D. & Potter, H. mRNA at the synapse: analysis of a synaptosomal preparation enriched in hippocampal dendritic spines. *The Journal of Neuroscience* **13**, 4054–4063 (1993).

138. Blüml, S. *et al.* Metabolic maturation of the human brain from birth through adolescence: insights from in vivo magnetic resonance spectroscopy. *Cereb Cortex* (2013) doi:10.1093/cercor/bhs283.

139. Petanjek, Z., Judaš, M., Kostović, I. & Uylings, H. B. M. Lifespan alterations of basal dendritic trees of pyramidal neurons in the human prefrontal cortex: A layer-specific pattern. *Cerebral Cortex* (2008) doi:10.1093/cercor/bhm124.

140. Somel, M. *et al.* MicroRNA-Driven Developmental Remodeling in the Brain Distinguishes Humans from Other Primates. *PLoS Biol* **9**, e1001214 (2011).

141. Song, L. *et al.* STAB: a spatio-temporal cell atlas of the human brain. *Nucleic Acids Res* **49**, D1029–D1037 (2021).

142. Lipovich, L. *et al.* Developmental changes in the transcriptome of human cerebral cortex tissue: Long noncoding RNA transcripts. *Cerebral Cortex* (2014) doi:10.1093/cercor/bhs414.

143. Zhang, X. Q., Wang, Z. L., Poon, M. W. & Yang, J. H. Spatial-temporal transcriptional dynamics of long non-coding RNAs in human brain. *Hum Mol Genet* (2017) doi:10.1093/hmg/ddx203.

144. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* (2008) doi:10.1073/pnas.0706729105.

145. Jarroux, J., Morillon, A. & Pinskaya, M. History, discovery, and classification of lncRNAs. in *Advances in Experimental Medicine and Biology* (2017). doi:10.1007/978-981-10-5203-3_1.

146. Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* (2015) doi:10.1038/ncomms8743.

147. Faghihi, M. A. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nat Med* (2008) doi:10.1038/nm1784.

148. Ramos, A. D. *et al.* The long noncoding RNA Pnky regulates neuronal differentiation of embryonic and postnatal neural stem cells. *Cell Stem Cell* (2015) doi:10.1016/j.stem.2015.02.007.

149. Ng, S. Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol Cell* (2013) doi:10.1016/j.molcel.2013.07.017.

150. Choi, S.-W., Kim, H.-W. & Nam, J.-W. The small peptide world in long noncoding RNAs. *Brief Bioinform* **20**, 1853–1864 (2019).

151. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* (2007) doi:10.1101/gr.6036807.

152. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (2012) doi:10.1101/gr.132159.111.

153. Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* (2006) doi:10.1101/gr.4200206.

154. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* (2012) doi:10.1038/nature11233.

155. Liu, S. J. *et al.* Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* (2016) doi:10.1186/s13059-016-0932-1.

156. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* (2013) doi:10.7554/eLife.01749.

157. Lin, N. *et al.* An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol Cell* (2014) doi:10.1016/j.molcel.2014.01.021.

158. Chalei, V. *et al.* The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. *Elife* (2014) doi:10.7554/eLife.04530.

159. Modarresi, F. *et al.* Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat Biotechnol* (2012) doi:10.1038/nbt.2158.

160. Tan, M. C. *et al.* The activity-induced long non-coding RNA Meg3 modulates AMPA receptor surface expression in primary cortical neurons. *Front Cell Neurosci* (2017) doi:10.3389/fncel.2017.00124.

161. Morikawa, T. & Manabe, T. Aberrant regulation of alternative pre-mRNA splicing in schizophrenia. *Neurochemistry International* Preprint at https://doi.org/10.1016/j.neuint.2010.08.012 (2010).

162. Kerin, T. *et al.* A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci Transl Med* (2012) doi:10.1126/scitranslmed.3003479.

163. Jang, Y. *et al.* Dysregulated long non-coding RNAs in the temporal lobe epilepsy mouse model. *Seizure* (2018) doi:10.1016/j.seizure.2018.04.010.

164. Zimmer-Bensch, G. Emerging Roles of Long Non-Coding RNAs as Drivers of Brain Evolution. *Cells* Preprint at https://doi.org/10.3390/cells8111399 (2019).

165. Kleaveland, B., Shi, C. Y., Stefano, J. & Bartel, D. P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* (2018) doi:10.1016/j.cell.2018.05.022.

166. Oliver, P. L. *et al.* Disruption of Visc-2, a brain-expressed conserved long noncoding RNA, does not elicit an overt anatomical or behavioral phenotype. *Cerebral Cortex* (2015) doi:10.1093/cercor/bhu196.

167. Han, X. *et al.* Mouse knockout models reveal largely dispensable but context-dependent functions of lncRNAs during development. *Journal of Molecular Cell Biology* Preprint at https://doi.org/10.1093/jmcb/mjy003 (2018).

168. Eissmann, M. *et al.* Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biol* **9**, 1076–1087 (2012).

169. Gao, F., Cai, Y., Kapranov, P. & Xu, D. Reverse-genetics studies of lncRNAs—what we have learnt and paths forward. *Genome Biol* **21**, 93 (2020).

170. Lee, H., Zhang, Z. & Krause, H. M. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? *Trends in Genetics* **35**, 892–902 (2019).

171. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (1979)* **355**, (2017).

172. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9–14 (2021).

173. Narayanan, R. T., Udvary, D. & Oberlaender, M. Cell Type-Specific Structural Organization of the Six Layers in Rat Barrel Cortex. *Front Neuroanat* **11**, (2017).

174. Radnikow, G. & Feldmeyer, D. Layer- and Cell Type-Specific Modulation of Excitatory Neuronal Activity in the Neocortex. *Front Neuroanat* **12**, (2018).

175. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. *BioEssays* **42**, 1900221 (2020).

176. 10X Genomics. Visium Spatial Gene Expression Reagent Kits User Guide, Document Number CG000239 Rev D. https://assets.ctfassets.net/an68im79xiti/3GGIfH3RWpd1bFVha1pexR/8baa08d9007157592b65b2cdc7130990/CG000239_VisiumSpatialGeneExpression_UserGuide_RevD.pdf.

177. Saiselet, M. *et al.* Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *J Mol Cell Biol* **12**, 906–908 (2021).

178. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology 2022 40:5* **40**, 661–671 (2022).

179. Zeng, H. *et al.* Large-Scale Cellular-Resolution Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures. *Cell* **149**, 483–496 (2012).

180. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

181. He, Z. *et al.* Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nat Neurosci* **20**, 886–895 (2017).

182. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**, e1006245 (2018).

183. Nayak, R. & Hasija, Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* **113**, 606–619 (2021).

184. Pasquini, G., Rojo Arias, J. E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* **19**, 961–969 (2021).

185. Chen, W. *et al.* A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Comput Struct Biotechnol J* **18**, 861–873 (2020).

186. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).

187. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255–261 (2018).

188. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* (2019) doi:10.1186/s12859-019-2599-6.

189. Mou, T., Deng, W., Gu, F., Pawitan, Y. & Vu, T. N. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Front Genet* **10**, (2020).

190. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20**, 269 (2019).

191. Xiang, R. *et al.* A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front Genet* **12**, (2021).

192. Xi, N. M. & Li, J. J. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Syst* **12**, 176-194.e6 (2021).

193. Kim, T. *et al.* Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform* **20**, 2316–2326 (2019).

194. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

195. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* (2019) doi:10.1016/j.cels.2018.11.005.

196. 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.e4 (2019).

197. DePasquale, E. A. K. *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep* **29**, 1718-1727.e8 (2019).

198. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, (2020).

199. Harvard Chan Bioinformatics Core. Introduction to single-cell RNA-seq. https://hbctraining.github.io/scRNA-seq_online/ (2020).

200. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* (2015) doi:10.1038/nbt.3192.

201. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).

202. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* (2019) doi:10.1186/s13059-019-1874-1.

203. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2**, (2020).

204. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

205. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019) doi:10.1016/j.cell.2019.05.031.

206. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, (2019).

207. Welch, J. *et al.* Integrative inference of brain cell similarities and differences from single-cell genomics. *bioRxiv* Preprint at (2018).

208. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* **21**, (2020).

209. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* (2014) doi:10.1186/s13059-014-0550-8.

210. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* **38**, 147–150 (2020).

211. Cao, Y., Kitanovski, S., Küppers, R. & Hoffmann, D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nat Biotechnol* **39**, 158–159 (2021).

212. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* Preprint at (2018).

213. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur Phys J B* **86**, 471 (2013).

214. Innes, B. T. & Bader, G. D. scClustViz – Single-cell RNAseq cluster assessment and visualization. *F1000Res* **7**, (2019).

215. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *bioRxiv* 2022.10.12.511898 (2022) doi:10.1101/2022.10.12.511898.

216. Clarke, Z. A. *et al.* Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols* vol. 16 Preprint at https://doi.org/10.1038/s41596-021-00534-0 (2021).

217. Shao, X. *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* (2020) doi:10.1016/j.isci.2020.100882.

218. Cao, Y., Wang, X. & Peng, G. SCSA: A cell type annotation tool for single-cell RNA-seq data. *Front Genet* (2020) doi:10.3389/fgene.2020.00490.

219. 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).

220. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun* **12**, 738 (2021).

221. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278 (2015).

222. Zhang, M. *et al.* IDEAS: individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol* **23**, 33 (2022).

223. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, (2014).

224. Tran, T. N. & Bader, G. D. Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data. *PLoS Comput Biol* **16**, (2020).

225. Street, K. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, (2018).

226. Macnair, W., Gupta, R. & Claassen, M. psupertime: supervised pseudotime analysis for time-series single-cell RNA-seq data. *Bioinformatics* **38**, i290–i298 (2022).

227. Blake, J. A. *et al.* Gene Ontology annotations and resources. *Nucleic Acids Res* **41**, (2013).

228. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols 2019 14:2* **14**, 482–517 (2019).

229. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* (2005) doi:10.1073/pnas.0506580102.

230. 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 (2019).

231. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–W97 (2016).

232. Bergen, D. C. Preventable Neurological Diseases Worldwide. *Neuroepidemiology* **17**, 67–73 (1998).

233. Aarli, J. A., Dua, T., Janca, A. & Muscettaorld, A. *Neurological disorders : public health challenges*. (2006).

234. Schutte, C. M. Analysis of HIV-related mortality data in a tertiary South African neurology unit, 2006- 2012. *South Afr J HIV Med* **14**, 121–124 (2013).

235. Rohlwink, U. K. *et al.* Clinical characteristics and neurodevelopmental outcomes of children with tuberculous meningitis and hydrocephalus. *Dev Med Child Neurol* **58**, 461–468 (2016).

236. Madhi, S. A. *et al.* High burden of invasive Streptococcus agalactiae disease in South African infants. *Ann Trop Paediatr* **23**, 15–23 (2003).

237. Ackermann, S., Le Roux, S. & Wilmshurst, J. M. Epidemiology of children with epilepsy at a tertiary referral centre in South Africa. *Seizure* **70**, 82–89 (2019).

238. Bell, S. H. *et al.* The Remarkably High Prevalence of Epilepsy and Seizure History in Fetal Alcohol Spectrum Disorders. *Alcohol Clin Exp Res* **34**, 1084–1089 (2010).

239. Chudley, A. E. Fetal Alcohol Spectrum Disorder—High Rates, High Needs, High Time for Action. *JAMA Pediatr* **171**, 940 (2017).

240. Oskoui, M., Coutinho, F., Dykeman, J., Jetté, N. & Pringsheim, T. An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. *Dev Med Child Neurol* **55**, 509–519 (2013).

241. Donald, K. A., Samia, P., Kakooza-Mwesige, A. & Bearden, D. Pediatric Cerebral Palsy in Africa: A Systematic Review. *Semin Pediatr Neurol* **21**, 30–35 (2014).

242. The African Child Policy Forum. *Children with disabilities in South Africa: The hidden reality*. (2011).

243. Laughton, B., Cornell, M., Boivin, M. & Van Rie, A. Neurodevelopment in perinatally HIV-infected children: a concern for adolescence. *J Int AIDS Soc* **16**, 18603 (2013).

244. Rohlwink, U. K. *et al.* Tuberculous meningitis in children is characterized by compartmentalized immune responses and neural excitotoxicity. *Nat Commun* **10**, 3767 (2019).

245. van Eyk, C. L. *et al.* Analysis of 182 cerebral palsy transcriptomes points to dysregulation of trophic signalling pathways and overlap with autism. *Transl Psychiatry* **8**, 88 (2018).

246. Farhadian, S. F. *et al.* Single-cell RNA sequencing reveals microglia-like cells in cerebrospinal fluid during virologically suppressed HIV. *JCI Insight* **3**, (2018).

247. Manyelo, C. M., Solomons, R. S., Walzl, G. & Chegou, N. N. Tuberculous Meningitis: Pathogenesis, Immune Responses, Diagnostic Challenges, and the Potential of Biomarker-Based Approaches. *J Clin Microbiol* **59**, (2021).

248. Shi, A. C., Rohlwink, U., Scafidi, S. & Kannan, S. Microglial Metabolism After Pediatric Traumatic Brain Injury – Overlooked Bystanders or Active Participants? *Front Neurol* **11**, (2021).

249. Schwab, N. *et al.* Neurons and glial cells acquire a senescent signature after repeated mild traumatic brain injury in a sex-dependent manner. *Front Neurosci* **16**, (2022).

250. Zhang, Z. *et al.* Systemic dendrimer-drug nanomedicines for long-term treatment of mild-moderate cerebral palsy in a rabbit model. *J Neuroinflammation* **17**, 319 (2020).

251. Borrajo, A., Spuch, C., Penedo, M. A., Olivares, J. M. & Agís-Balboa, R. C. Important role of microglia in HIV-1 associated neurocognitive disorders and the molecular pathways implicated in its pathogenesis. *Ann Med* **53**, 43–69 (2021).

252. Farhadian, S. F. *et al.* Single-cell RNA sequencing reveals microglia-like cells in cerebrospinal fluid during virologically suppressed HIV. *JCI Insight* **3**, (2018).

253. Tripathi, S. *et al.* Glial alterations in tuberculous and cryptococcal meningitis and their relation to HIV co-infection – A study on human brains. *The Journal of Infection in Developing Countries* **8**, 1421–1443 (2014).

254. O'Malley, J., Wardlaw, T., You, D., Hug, L. & Anthony, D. Africa's child demographics and the world's future. *The Lancet* vol. 384 Preprint at https://doi.org/10.1016/S0140-6736(14)61331-3 (2014).

255. Thrupp, N. *et al.* Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Rep* **32**, 108189 (2020).

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

257. 10X Genomics. Nuclei Isolation from Cell Suspensions & Tissues for Single Cell RNA Sequencing (CG000124, Rev E). https://assets.ctfassets.net/an68im79xiti/6x4KMzpIgPgkje01sR1Xgr/9cfb7d859985e5c479aec4e0e501f903/CG000124_Demonstrated_Protocol_Nuclei_isolation_RevE.pdf (2021).

258. Pantano, L. DEGreport: Report of DEG analysis. Preprint at https://doi.org/10.18129/B9.bioc.DEGreport (2017).

259. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* (2009) doi:10.1038/nature07672.

260. Sanchez-Taltavull, D. *et al.* Bayesian correlation is a robust gene similarity measure for single-cell RNA-seq data. *NAR Genom Bioinform* **2**, (2020).

261. Lin, J. *et al.* Pipelines for cross-species and genome-wide prediction of long noncoding RNA binding. *Nature Protocols 2019 14:3* **14**, 795–818 (2019).

262. Wen, Y., Wu, Y., Xu, B., Lin, J. & Zhu, H. Fasim-LongTarget enables fast and accurate genome-wide lncRNA/DNA binding prediction. *Comput Struct Biotechnol J* **20**, 3347–3350 (2022).

263. Duca, M., Vekhoff, P., Oussedik, K., Halby, L. & Arimondo, P. B. The triple helix: 50 years later, the outcome. *Nucleic Acids Res* **36**, 5123–5138 (2008).

264. Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).

265. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667–11672 (2009).

266. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Preprint at (2016).

267. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

268. Tippani, M. *et al.* VistoSeg: processing utilities for high-resolution Visium/Visium-IF images for spatial transcriptomics data. *bioRxiv* (2021) doi:https://doi.org/10.1101/2021.08.04.452489.

269. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517–526 (2022).

270. Patir, A., Shih, B., McColl, B. W. & Freeman, T. C. A core transcriptional signature of human microglia: Derivation and utility in describing region-dependent alterations associated with Alzheimer's disease. *Glia* **67**, 1240–1253 (2019).

271. Zheng, K. *et al.* Molecular and Genetic Evidence for the PDGFRα-Independent Population of Oligodendrocyte Progenitor Cells in the Developing Mouse Brain. *The Journal of Neuroscience* **38**, 9505 (2018).

272. Eng, L. F. Glial fibrillary acidic protein (GFAP): the major protein of glial intermediate filaments in differentiated astrocytes. *J Neuroimmunol* **8**, 203–214 (1985).

273. Breuer, M. *et al.* QDPR homologues in Danio rerio regulate melanin synthesis, early gliogenesis, and glutamine homeostasis. *PLoS One* **14**, (2019).

274. Huang, Z. *et al.* Presynaptic HCN1 channels regulate Ca V 3.2 activity and neurotransmission at select cortical synapses. **14**, (2011).

275. Mantegazza, M., Cestèle, S. & Catterall, W. A. Sodium channelopathies of skeletal muscle and brain. *Physiol Rev* **101**, 1633–1689 (2021).

276. Jules Gilet. Quality Control - sequencing and mapping | Single RNA-seq data analysis with R. https://nbisweden.github.io/excelerate-scRNAseq/session-seqmap/sequencing_qc.html.

277. Bloom, J. D. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ* **2018**, (2018).

278. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* (2009) doi:10.1371/journal.pcbi.1000598.

279. Lake, B. B. *et al.* A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nat Commun* (2019) doi:10.1038/s41467-019-10861-2.

280. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2019) doi:10.1038/nbt.4314.

281. Khatri, R. & Bonn, S. Uncertainty Estimation for Single-cell Label Transfer. *Proc Mach Learn Res* **179**, 109–128 (2022).

282. Takara Bio. SMART-Seq V4 Ultra Low Input RNA Kit for sequencing.

283. Johansen, N. & Quon, G. ScAlign: A tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol* **20**, (2019).

284. Batiuk, M. Y. *et al.* Upper cortical layer–driven network impairment in schizophrenia. *Sci Adv* **8**, (2022).

285. Petukhov, V. *et al.* Case-control analysis of single-cell RNA-seq studies. *bioRxiv* (2022) doi:10.1101/2022.03.15.484475.

286. A qualitative and quantitative ultrastructural study of glial cells in the developing visual cortex of the rat. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **301**, 55–84 (1983).

287. Sorrell, F. J. *et al.* STK32A is a dual-specificity AGC kinase with a preference for acidic substrates. *bioRxiv* 2020.03.04.976555 (2020) doi:10.1101/2020.03.04.976555.

288. Alkaslasi, M. R. *et al.* Single nucleus RNA-sequencing defines unexpected diversity of cholinergic neuron types in the adult mouse spinal cord. *Nat Commun* **12**, (2021).

289. Pérez-Palma, E. *et al.* Early Transcriptional Changes Induced by Wnt/β-Catenin Signaling in Hippocampal Neurons. *Neural Plast* **2016**, (2016).

290. Lie, D. C. *et al.* Wnt signalling regulates adult hippocampal neurogenesis. *Nature* **437**, 1370–1375 (2005).

291. Chen, J., Chang, S. P. & Tang, S. J. Activity-dependent synaptic Wnt release regulates hippocampal long term potentiation. *J Biol Chem* **281**, 11910–11916 (2006).

292. Mata, A. *et al.* New functions of Semaphorin 3E and its receptor PlexinD1 during developing and adult hippocampal formation. *Scientific Reports 2018 8:1* **8**, 1–16 (2018).

293. Chauvet, S. *et al.* Gating of Sema3E/PlexinD1 Signaling by Neuropilin-1 Switches Axonal Repulsion to Attraction during Brain Development. *Neuron* **56**, 807 (2007).

294. Keller, D., Tsuda, M. C., Usdin, T. B. & Dobolyi, A. Behavioural actions of tuberoinfundibular peptide 39 (parathyroid hormone 2). *J Neuroendocrinol* **34**, e13130 (2022).

295. Wang, K. C. & Chang, H. Y. Molecular Mechanisms of Long Noncoding RNAs. *Mol Cell* **43**, 904–914 (2011).

296. Murphy, A. E. *et al.* A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nature Communications 2022 13:1* **13**, 1–4 (2022).

297. Heidel, R. E. Causality in Statistical Power: Isomorphic Properties of Measurement, Research Design, Effect Size, and Sample Size. *Scientifica* vol. 2016 Preprint at https://doi.org/10.1155/2016/8920418 (2016).

298. Hezroni, H. *et al.* Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep* **11**, 1110 (2015).

299. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, (2017).

300. López-Bendito, G. *et al.* Robo1 and Robo2 cooperate to control the guidance of major axonal tracts in the mammalian forebrain. *Journal of Neuroscience* **27**, (2007).

301. Blockus, H. *et al.* Synaptogenic activity of the axon guidance molecule Robo2 underlies hippocampal circuit function. *Cell Rep* **37**, 109828 (2021).

302. Schreiner, D., Savas, J. N., Herzog, E., Brose, N. & de Wit, J. Synapse biology in the 'circuit-age'-paths toward molecular connectomics. *Curr Opin Neurobiol* **42**, 102–110 (2017).

303. Simchovitz, A. *et al.* A lncRNA survey finds increases in neuroprotective LINC-PINT in Parkinson's disease substantia nigra. *Aging Cell* **19**, (2020).

304. Von Schimmelmann, M. *et al.* Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration. *Nat Neurosci* **19**, 1321–1330 (2016).

305. Zhang, M. *et al.* A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat Commun* **9**, 4475 (2018).

306. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2006-2010. *Neuro Oncol* **15**, ii1–ii56 (2013).

307. Zhang, X. *et al.* BAIAP3, a C2 domain–containing Munc13 protein, controls the fate of dense-core vesicles in neuroendocrine cells. *J Cell Biol* **216**, 2151 (2017).

308. Kim, H. *et al.* Baiap3 regulates depressive behaviors in mice via attenuating dense core vesicle trafficking in subsets of prefrontal cortex neurons. *Neurobiol Stress* **16**, 100423 (2022).

309. Yokoyama, K. *et al.* NYAP: a phosphoprotein family that links PI3K to WAVE1 signalling in neurons. *EMBO J* **30**, 4739–4754 (2011).

310. Ho, H. Y. H. *et al.* Toca-1 mediates Cdc42-dependent actin nucleation by activating the N-WASP-WIP complex. *Cell* **118**, 203–216 (2004).

311. Rodriguez-Murillo, L. *et al.* Fine mapping on chromosome 13q32-34 and brain expression analysis implicates MYO16 in schizophrenia. *Neuropsychopharmacology* **39**, 934–943 (2014).

312. 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).

313. Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry 2014 19:2* **19**, 253–258 (2013).

314. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry 2011 16:10* **16**, 996–1005 (2011).

315. Prokopenko, D. *et al.* Whole-genome sequencing reveals new Alzheimer's disease–associated rare variants in loci related to synaptic function and neuronal development. *Alzheimer's and Dementia* **17**, (2021).

316. Strittmatter, W. J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A* **90**, 1977 (1993).

317. Hanley, J. G. The regulation of AMPA receptor endocytosis by dynamic protein-protein interactions. *Front Cell Neurosci* **12**, (2018).

318. Jackson, J. *et al.* Targeting the synapse in Alzheimer's disease. *Frontiers in Neuroscience* vol. 13 Preprint at https://doi.org/10.3389/fnins.2019.00735 (2019).

319. Zalocusky, K. A. *et al.* Neuronal ApoE upregulates MHC-I expression to drive selective neurodegeneration in Alzheimer's disease. *Nature Neuroscience 2021 24:6* **24**, 786–798 (2021).

320. Arias-Vásquez, A. *et al.* A Potential Role for the STXBP5-AS1 Gene in Adult ADHD Symptoms. *Behav Genet* **49**, 270–285 (2019).

321. Cupertino, R. B. *et al.* SNARE complex in developmental psychiatry: neurotransmitter exocytosis and beyond. *Journal of Neural Transmission 2016 123:8* **123**, 867–883 (2016).

322. Gillingham, A. K. & Munro, S. The Small G Proteins of the Arf Family and Their Regulators. *https://doi.org/10.1146/annurev.cellbio.23.090506.123209* **23**, 579–611 (2007).

323. He, L. *et al.* Exome-wide age-of-onset analysis reveals exonic variants in ERN1 and SPPL2C associated with Alzheimer's disease. *Transl Psychiatry* **11**, (2021).

324. Allen, M. *et al.* Gene expression, methylation and neuropathology correlations at progressive supranuclear palsy risk loci. *Acta Neuropathol* **132**, 197–211 (2016).

325. Jäkel, S. *et al.* Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature 2019 566:7745* **566**, 543–547 (2019).

326. Seeker, L. A. *et al.* Marked regional glial heterogeneity in the human white matter of the central nervous system. *bioRxiv* 2022.03.22.485367 (2022) doi:10.1101/2022.03.22.485367.

327. Klein, M. O. *et al.* Dopamine: Functions, Signaling, and Association with Neurological Diseases. *Cell Mol Neurobiol* **39**, 31–59 (2019).

328. Edgar, N. & Sibille, E. A putative functional role for oligodendrocytes in mood regulation. *Transl Psychiatry* **2**, e109 (2012).

329. Tanaka, M. *et al.* Adenosine A2B receptor down-regulates metabotropic glutamate receptor 5 in astrocytes during postnatal development. *Glia* **69**, (2021).

330. Avruch, J. *et al.* Insulin and amino-acid regulation of mTOR signaling and kinase activity through the Rheb GTPase. *Oncogene* vol. 25 Preprint at https://doi.org/10.1038/sj.onc.1209882 (2006).

331. Hall, M. N. mTOR—What Does It Do? *Transplant Proc* **40**, S5–S8 (2008).

332. Yamagatas, K. *et al.* rheb, a Growth Factor-and Synaptic Activity-regulated Gene, Encodes a Novel Ras-related Protein". *The Journal of biological chemistryGICAL CHEMISTRY* **269**, 16333–16339 (1994).

333. Keane, L. *et al.* mTOR-dependent translation amplifies microglia priming in aging mice. *Journal of Clinical Investigation* **131**, (2021).

334. Huber, L. A., Dupree, P. & Dotti, C. G. A deficiency of the small GTPase rab8 inhibits membrane traffic in developing neurons. *Mol Cell Biol* **15**, (1995).

335. Johari, A. H. *et al.* The rs1986112 Variant is Associated with Increased RAB8B Gene Expression in Schizophrenic Patients. *Clin Lab* **65**, (2019).

336. Monji, A., Kato, T. & Kanba, S. Cytokines and schizophrenia: Microglia hypothesis of schizophrenia. *Psychiatry Clin Neurosci* **63**, 257–265 (2009).

337. King, R. W., Jackson, P. K. & Kirschner, M. W. Mitosis in transition. *Cell* vol. 79 Preprint at https://doi.org/10.1016/0092-8674(94)90542-8 (1994).

338. Bandeira, F., Lent, R. & Herculano-Houzel, S. Changing numbers of neuronal and non-neuronal cells underlie postnatal brain growth in the rat. *Proc Natl Acad Sci U S A* **106**, (2009).

339. Mermer, F. *et al.* Astrocytic GABA transporter 1 deficit in novel SLC6A1 variants mediated epilepsy: Connected from protein destabilization to seizures in mice and humans. *Neurobiol Dis* **172**, (2022).

340. Wilfong, A., Nordli, D. R. & Dashe, J. F. Seizures and epilepsy in children: Classification, etiology, and clinical features - UpToDate. *UpToDate* (2022).

341. Longden, T. A. *et al.* Local IP 3 receptor–mediated Ca 2+ signals compound to direct blood flow in brain capillaries. *Sci Adv* **7**, (2021).

342. Wu, C. *et al.* Age-Related Changes of Normal Cerebral and Cardiac Blood Flow in Children and Adults Aged 7 Months to 61 Years. *J Am Heart Assoc* **5**, (2016).

343. Cooper, C. M. *et al.* Discovery and replication of cerebral blood flow differences in major depressive disorder. *Mol Psychiatry* **25**, 1500–1510 (2020).

344. Sonal Sekhar, M., Sasidharan, S., Joseph, S. & Kumar, A. Migraine management: How do the adult and paediatric migraines differ? *Saudi Pharmaceutical Journal : SPJ* **20**, 1 (2012).

345. Victor, T., Hu, X., Campbell, J., Buse, D. & Lipton, R. Migraine prevalence by age and sex in the United States: A life-span study. *Cephalalgia* **30**, 1065–1072 (2010).

346. Camfield, P. & Camfield, C. Incidence, prevalence and aetiology of seizures and epilepsy in children. *Epileptic Disord* **17**, 117–123 (2015).

347. Raol, Y. H., Lynch, D. R. & Brooks-Kayal, A. R. Role of excitatory amino acids in developmental epilepsies. *Ment Retard Dev Disabil Res Rev* **7**, 254–260 (2001).

348. Hu, C., Tao, L., Cao, X. & Chen, L. The solute carrier transporters and the brain: Physiological and pharmacological implications. *Asian J Pharm Sci* **15**, 131 (2020).

349. Salatino-Oliveira, A., Rohde, L. A. & Hutz, M. H. The dopamine transporter role in psychiatric phenotypes. *Am J Med Genet B Neuropsychiatr Genet* **177**, 211–231 (2018).

350. Tordera, R. M. *et al.* Enhanced anxiety, depressive-like behaviour and impaired recognition memory in mice with reduced expression of the vesicular glutamate transporter 1 (VGLUT1). *European Journal of Neuroscience* **25**, 281–290 (2007).

351. Moore, Y. E., Kelley, M. R., Brandon, N. J., Deeb, T. Z. & Moss, S. J. Seizing Control of KCC2: A New Therapeutic Target for Epilepsy. *Trends Neurosci* **40**, 555–571 (2017).

352. Rask-Andersen, M., Masuram, S., Fredriksson, R. & Schiöth, H. B. Solute carriers as drug targets: Current use, clinical trials and prospective. *Mol Aspects Med* **34**, 702–710 (2013).

353. Wilfong, A., Nordli, D. R. & Dashe, J. F. Seizures and epilepsy in children: Initial treatment and monitoring - UpToDate. *UpToDate* (2022).

354. Willem, M. *et al.* Control of peripheral nerve myelination by the β-secretase BACE1. *Science (1979)* **314**, 664–666 (2006).

355. Hitt, B. *et al.* β-Site amyloid precursor protein (APP)-cleaving enzyme 1 (BACE1)-deficient mice exhibit a close homolog of L1 (CHL1) loss-of-function phenotype involving axon guidance defects. *Journal of Biological Chemistry* **287**, 38408–38425 (2012).

356. Hu, X. *et al.* Bace1 modulates myelination in the central and peripheral nervous system. *Nature Neuroscience 2006 9:12* **9**, 1520–1525 (2006).

357. Lombardo, S. *et al.* BACE1 partial deletion induces synaptic plasticity deficit in adult mice. *Scientific Reports 2019 9:1* **9**, 1–14 (2019).

358. Harrington, A. J. *et al.* MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *Elife* **5**, (2016).

359. Zhou, W. Z. *et al.* Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype–phenotype correlations. *Hum Mutat* **40**, 801 (2019).

360. Mitchell, A. C. *et al.* MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice. *Mol Psychiatry* **23**, 123–132 (2018).

361. Yu, Q., Zhao, M. W. & Yang, P. LncRNA UCA1 Suppresses the Inflammation Via Modulating miR-203-Mediated Regulation of MEF2C/NF-κB Signaling Pathway in Epilepsy. *Neurochem Res* **45**, 783–795 (2020).

362. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet* **48**, 1031–1036 (2016).

363. Tsai, N. P. *et al.* Multiple autism-linked genes mediate synapse elimination via proteasomal degradation of a synaptic scaffold PSD-95. *Cell* **151**, 1581–1594 (2012).

364. Adachi, M., Lin, P. Y., Pranav, H. & Monteggia, L. M. Postnatal Loss of Mef2c Results in Dissociation of Effects on Synapse Number and Learning and Memory. *Biol Psychiatry* **80**, 140–148 (2016).

365. Kaupmann, K. *et al.* GABA(B)-receptor subtypes assemble into functional heteromeric complexes. *Nature* **396**, 683–7 (1998).

366. Gonchar, Y., Pang, L., Malitschek, B., Bettler, B. & Burkhalter, A. Subcellular localization of GABAB receptor subunits in rat visual cortex. *J Comp Neurol* **431**, 182–197 (2001).

367. Kuriyama, K., Hirouchi, M. & Kimura, H. Neurochemical and molecular pharmacological aspects of the GABA(B) receptor. *Neurochem Res* **25**, 1233–9 (2000).

368. Agner, C. GABA in the nervous system: The view at fifty years. *J Neurol Sci* **190**, 101 (2001).

369. Fukui, M. *et al.* Gradual downregulation of protein expression of the partner GABA(B)R2 subunit during postnatal brain development in mice defective of GABA(B)R1 subunit. *J Pharmacol Sci* **115**, 45–55 (2011).

370. Fatemi, S. H., Folsom, T. D., Reutiman, T. J. & Thuras, P. D. Expression of GABAB receptors is altered in brains of subjects with autism. *Cerebellum* **8**, 64 (2009).

371. Princivalle, A. P., Duncan, J. S., Thom, M. & Bowery, N. G. GABAB1a, GABAB1b and GABAB2 mRNA variants expression in hippocampus resected from patients with temporal lobe epilepsy. *Neuroscience* **122**, 975–984 (2003).

372. Fatemi, S. H., Folsom, T. D. & Thuras, P. D. Deficits in GABAB receptor system in schizophrenia and mood disorders: a postmortem study. *Schizophr Res* **128**, 37 (2011).

373. Chen, J., Yen, A., Florian, C. P. & Dougherty, J. D. MYT1L in the making: emerging insights on functions of a neurodevelopmental disorder gene. *Translational Psychiatry 2022 12:1* **12**, 1–8 (2022).

374. Chen, J. *et al.* A MYT1L syndrome mouse model recapitulates patient phenotypes and reveals altered brain development due to disrupted neuronal maturation. *Neuron* **109**, 3775 (2021).

375. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* **35**, 357–379 (2019).

376. Dulka, B. N., Pullins, S. E., Cullen, P. K., Moyer, J. R. & Helmstetter, F. J. Age-Related Memory Deficits are Associated with Changes in Protein Degradation in Brain Regions Critical for Trace Fear Conditioning. *Neurobiol Aging* **91**, 160 (2020).

377. Robinson, R. B. & Siegelbaum, S. A. Hyperpolarization-Activated Cation Currents: From Molecules to Physiological Function. *Annu Rev Physiol* **65**, 453–480 (2003).

378. Tsay, D., Dudman, J. T. & Siegelbaum, S. A. HCN1 Channels Constrain Synaptically Evoked Ca2+ Spikes in Distal Dendrites of CA1 Pyramidal Neurons. *Neuron* **56**, 1076 (2007).

379. Yang, S. S. *et al.* Cell-type specific development of the hyperpolarization-activated current, Ih, in prefrontal cortical neurons. *Front Synaptic Neurosci* **10**, 7 (2018).

380. Cheah, C. S. *et al.* Correlations in timing of sodium channel expression, epilepsy, and sudden death in Dravet syndrome. *Channels* **7**, (2013).

381. Heighway, J. *et al.* Sodium channel expression and transcript variation in the developing brain of human, Rhesus monkey, and mouse. *Neurobiol Dis* **164**, 105622 (2022).

382. Yu, F. H. *et al.* Reduced sodium current in GABAergic interneurons in a mouse model of severe myoclonic epilepsy in infancy. *Nat Neurosci* **9**, 1142–1149 (2006).

383. Nava, C. *et al.* De novo mutations in HCN1 cause early infantile epileptic encephalopathy. *Nat Genet* **46**, 640–645 (2014).

384. Deneen, B. *et al.* The Transcription Factor NFIA Controls the Onset of Gliogenesis in the Developing Spinal Cord. *Neuron* **52**, (2006).

385. Suzuki, N. *et al.* Teneurin-4 is a novel regulator of oligodendrocyte differentiation and myelination of small-diameter axons in the CNS. *J Neurosci* **32**, 11586–11599 (2012).

386. Hor, H. *et al.* Missense mutations in TENM4 , a regulator of axon guidance and central myelination, cause essential tremor. *Hum Mol Genet* **24**, 5677–5686 (2015).

387. 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. *Nat Commun* **13**, 2297 (2022).

388. Felix, L., Stephan, J. & Rose, C. R. Astrocytes of the early postnatal brain. *European Journal of Neuroscience* vol. 54 Preprint at https://doi.org/10.1111/ejn.14780 (2021).

389. Takano, T. *et al.* Chemico-genetic discovery of astrocytic control of inhibition in vivo. *Nature* **588**, (2020).

390. Kashiwabuchi, N. *et al.* Impairment of motor coordination, Purkinje cell synapse formation, and cerebellar long-term depression in GluRδ2 mutant mice. *Cell* **81**, 245–252 (1995).

391. Varoqueaux, F. *et al.* Neuroligins Determine Synapse Maturation and Function. *Neuron* **51**, 741–754 (2006).

392. Roman, D. L. & Traynor, J. R. Regulators of G Protein Signaling (RGS) Proteins as Drug Targets: Modulating G-Protein-Coupled Receptor (GPCR) Signal Transduction. *J Med Chem* **54**, 7433–7440 (2011).

393. Kofuji, P. & Araque, A. G-Protein-Coupled Receptors in Astrocyte–Neuron Communication. *Neuroscience* **456**, 71–84 (2021).

394. Mader, S. & Brimberg, L. Aquaporin-4 Water Channel in the Brain and Its Implication for Health and Disease. *Cells* **8**, 90 (2019).

395. Soomro, S. H., Jie, J. & Fu, H. Oligodendrocytes Development and Wnt Signaling Pathway. *International Journal of Human Anatomy* **1**, 17–35 (2018).

396. Lee, I.-H., Koelliker, E. & Kong, S. W. Endophenotype-Wide Association Study Reveals Genetic Substrates of Core Symptom Domains and Neurocognitive Function in Autism. *Res Sq* (2021) doi:https://doi.org/10.21203/rs.3.rs-948337/v1.

397. Bhalala, O. G., Nath, A. P., Inouye, M. & Sibley, C. R. Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. *PLoS Genet* **14**, (2018).

398. Li, Q. *et al.* Genome-wide association study of paliperidone efficacy. *Pharmacogenet Genomics* **27**, (2017).

399. Brunklaus, A. *et al.* The gain of function SCN1A disorder spectrum: novel epilepsy phenotypes and therapeutic implications. *Brain* **11**, 3816–3831 (2022).

400. Bender, A. C., Morse, R. P., Scott, R. C., Holmes, G. L. & Lenck-Santini, P. P. SCN1A mutations in Dravet syndrome: Impact of interneuron dysfunction on neural networks and cognitive outcome. *Epilepsy and Behavior* vol. 23 Preprint at https://doi.org/10.1016/j.yebeh.2011.11.022 (2012).

401. Boscia, F., Elkjaer, M. L., Illes, Z. & Kukley, M. Altered Expression of Ion Channels in White Matter Lesions of Progressive Multiple Sclerosis: What Do We Know About Their Function? *Frontiers in Cellular Neuroscience* vol. 15 Preprint at https://doi.org/10.3389/fncel.2021.685703 (2021).

402. Jensen, K. B., Musunuru, K., Lewis, H. A., Burley, S. K. & Darnell, R. B. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc Natl Acad Sci U S A* **97**, (2000).

403. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**, (2021).

404. Issler, O. *et al.* Sex-Specific Role for the Long Non-coding RNA LINC00473 in Depression. *Neuron* **106**, (2020).

405. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: Multiplexed, quantitative, sensitive, versatile, robust. *Development (Cambridge)* (2018) doi:10.1242/dev.165753.

406. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine 2020 26:5* **26**, 792–802 (2020).

407. Albert, P. R. Why is depression more prevalent in women? *Journal of Psychiatry and Neuroscience* vol. 40 Preprint at https://doi.org/10.1503/jpn.150205 (2015).

408. McLean, C. P., Asnaani, A., Litz, B. T. & Hofmann, S. G. Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. *J Psychiatr Res* **45**, (2011).

409. Chan Zuckerberg Initiative. Pediatric Networks for the Human Cell Atlas. https://chanzuckerberg.com/science/programs-resources/single-cell-biology/pediatric-networks/.

410. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun* **9**, (2018).

411. Barton, A. J. L., Pearson, R. C. A., Najlerahim, A. & Harrison, P. J. Pre-and Postmortem Influences on Brain RNA. *Journal of Neurochemistry* Preprint at https://doi.org/10.1111/j.1471-4159.1993.tb03532.x (1993).

412. Zhou, B. *et al.* GRID-seq for comprehensive analysis of global RNA–chromatin interactions. *Nat Protoc* (2019) doi:10.1038/s41596-019-0172-4.

413. McHugh, C. A. & Guttman, M. RAP-MS: A method to identify proteins that interact directly with a specific RNA molecule in cells. in *Methods in Molecular Biology* vol. 1649 (2018).

414. Clark, S. J. *et al.* Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* **12**, (2017).

415. Feng, H. *et al.* Complexity and graded regulation of neuronal cell-type–specific alternative splicing revealed by single-cell RNA sequencing. *Proceedings of the National Academy of Sciences* **118**, (2021).

416. Zhang, X. *et al.* Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell* **166**, 1147-1162.e15 (2016).

417. Hagemann-Jensen, M. *et al.* Smart-seq3 Protocol V.3. https://www.protocols.io/view/smart-seq3-protocol-bcq4ivyw?step=2 (2020).