Cross-Regional and Cross-Species Analysis of Gene Expression Changes at the Perinatal Boundary in Human Brain Development

Connor McPhan

Supervised by Chuck Herring, Ryan Lister

Formatting guide: [Bioinformatics](https://academic.oup.com/bioinformatics/issue)

This paper is submitted in partial fulfilment of the requirements for a

Master of Bioinformatics

SCIE5574 Masters Research Dissertation

UWA School of Agriculture and Environment

The University of Western Australia

May 2025

Actual word count: 5251

**Abstract**

Sweeping transcriptomic changes occur around birth in human brain development, with similar changes evident in mice at eye-opening. Although this period is likely to be critical for normal brain development, the changes that occur there are not well characterized. We used single-nucleus RNA-seq datasets from prior studies to investigate the gene expression changes underlying this period, comparing them across different regions of the brain, different species and different cell types. We find that the expression changes occurring at birth in humans, mice and opossums are similar and may extend through to eye-opening in the rodent species. Further, we present evidence for the possible involvement of Kruppel-like factors and Sp-like factors in regulating these changes and put forward candidates for the genes driving them based on conserved patterns of expression.

**Introduction**

Human brain development is an extraordinarily complex process underpinned by ongoing spatiotemporal changes in gene expression which progress in a uniquely protracted manner, extending well into adulthood (Li *et al.* 2020; Herring *et al.* 2022). Previous studies examining changes in gene expression over the course of development have consistently described a period of rapid change in the brain occurring around birth in humans and non-human primates, extending across brain regions, cell types and sub-types, and linked to the transition from the prenatal to postnatal environment (Bakken *et al.* 2016; Li *et al.* 2018; Zhu *et al.* 2018; Herring *et al.* 2022). Similar changes have been evident in mice at around 14 days following birth (Tan *et al.* 2021), coinciding with eye-opening. Given the extent and timing of these changes this period is a potentially crucial turning point in development and dysregulation at this time may be a driving factor for early onset neuropsychiatric disorders (Li *et al.* 2018; Herring *et al.* 2022). Improving understanding of the changes happening in this period could enable advances in neuropsychiatric disease research and organoid development. Many major neuropsychiatric diseases still have unclear genetic and molecular mechanisms and are still undergoing active research to elucidate these mechanisms and enable new avenues for treatment (Czarny *et al.* 2020; Joshi and Salton 2022; Repetto *et al.* 2024; Bendl *et al.* 2025; Huang *et al.* 2025). Cerebral organoids are cultured models of the human brain which are widely used in research to model neurological disease states and a variety of other developmental aspects (Jeong, Choi and Cho 2023; Kim and Chang 2023; Mulder *et al.* 2023) but are known to stall during differentiation and remain in a state resembling the fetal brain (Camp *et al.* 2015; Luo *et al.* 2016), limiting their usefulness as models of the adult brain. The regulatory mechanisms underlying perinatal changes in gene expression could be key pathways involved in the development of psychiatric disease when disrupted and be applied to organoid development to push differentiation past the fetal stage and into adolescent and adult states. Despite its research potential this period remains largely uncharacterized, with studies of this period tending to focus more on changes in disease genes (Li *et al.* 2018), tracing evolutionary pathways (Sepp *et al.* 2024) or provisioning new resources for exploration in further studies (Bakken *et al.* 2016). This study aims to fill this gap in two ways. First, by more fully characterizing the rapid changes in gene expression in the human brain around birth, which we hypothesise to be driven by some switch in functional focus, e.g. from tissue development to a response to stimulus. Second, by comparing these changes in different regions of the brain and in different species to see if the mechanisms and developmental timing are conserved, which may reveal highly conserved pathways critical for normal brain development.

**Results**

Data Preprocessing and Developmental Staging

We used publicly available single-nucleus RNA-seq data collected from two recent, peer-reviewed papers (Herring *et al.* 2022; Sepp *et al.* 2024), hereafter referred to as the Herring data and Sepp data, respectively. These datasets were chosen because they use similar 10x Genomics protocols for sequencing, span a similar wide range of the developmental course, including our time points of interest across multiple species, and were taken from different regions of the brain (prefrontal cortex and cerebellum, respectively), making them good candidates to facilitate investigations relevant to our hypothesis. After preprocessing, the data consisted of 447,556 total single nuclei and 103,204 total gene features which were used for analysis. We first sought to verify when the largest changes in expression were occurring for each dataset to ensure we would be able to capture these transitions in our analysis. Sample-to-sample distances were calculated for each dataset after transforming them with a variance stabilizing transformation and visualised using heatmaps (Figure 1), as recommended for sample clustering using the DESeq2 workflow (Anders and Huber 2010). This confirmed that in the human and mouse datasets significant changes occurred shortly before birth, in the late gestational period. The lack of fetal samples in the opossum data obscured a similar analysis there, however there did appear to be a transition from early to late neonatal periods between postnatal days 1 and 4, though this may be due to a lack of samples at the intervening ages.

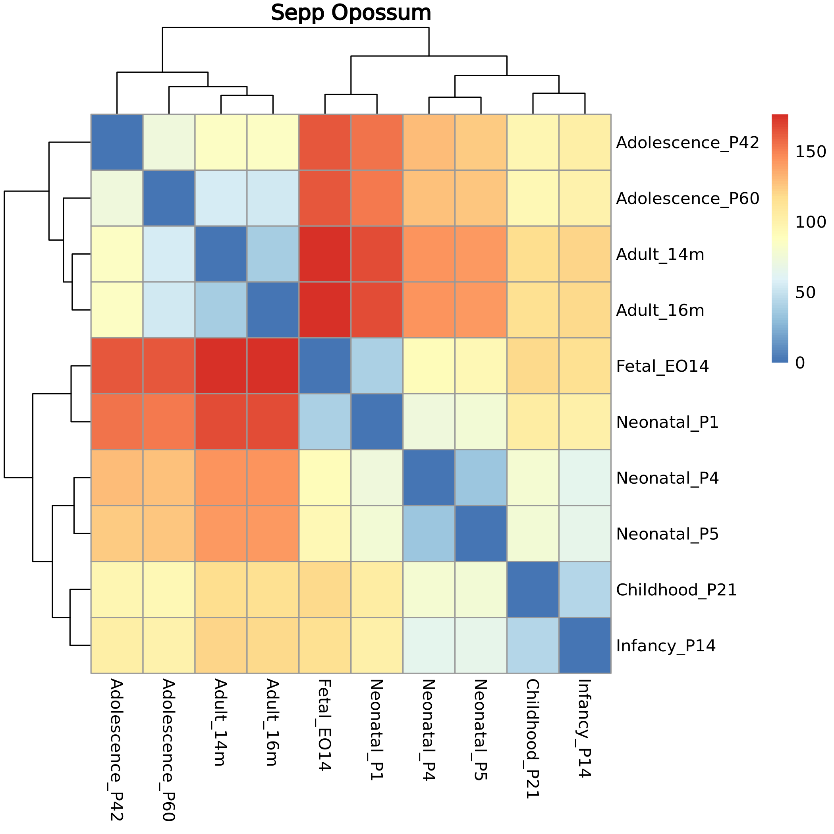
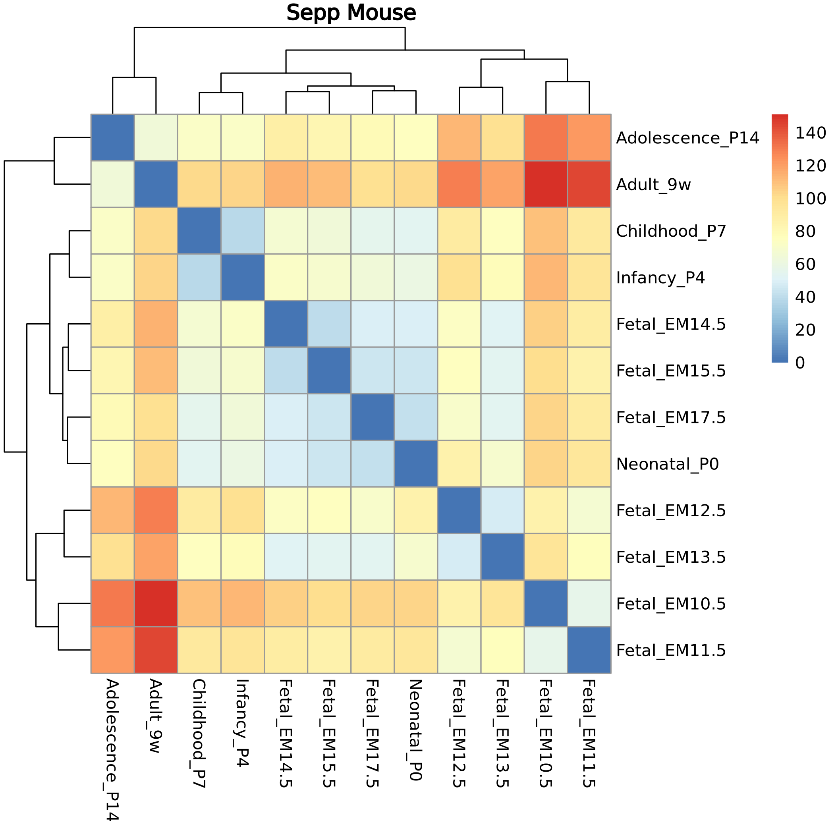
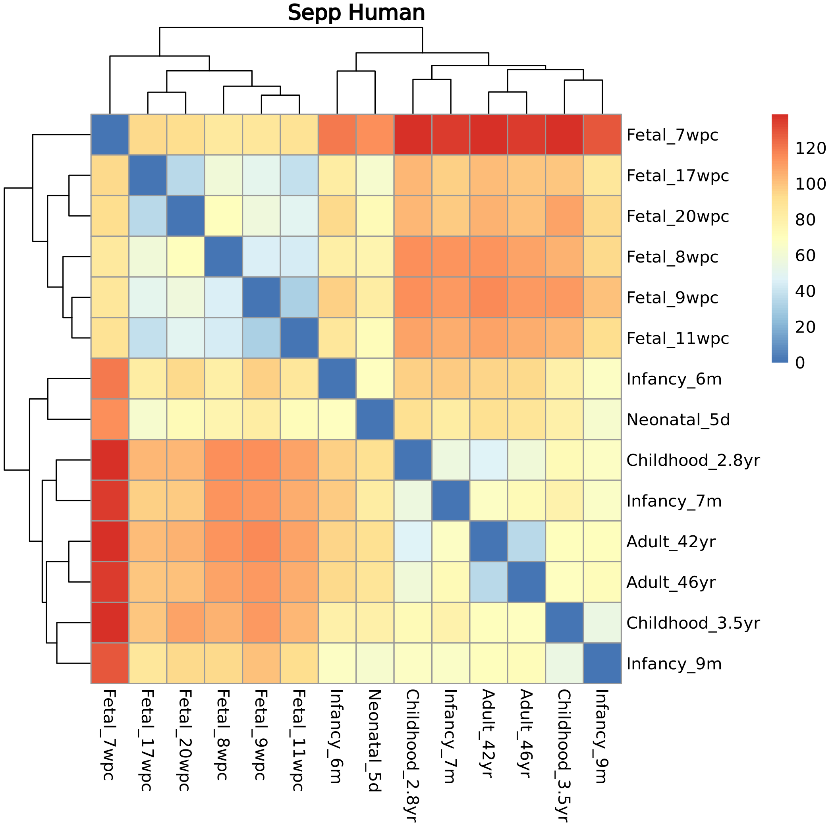
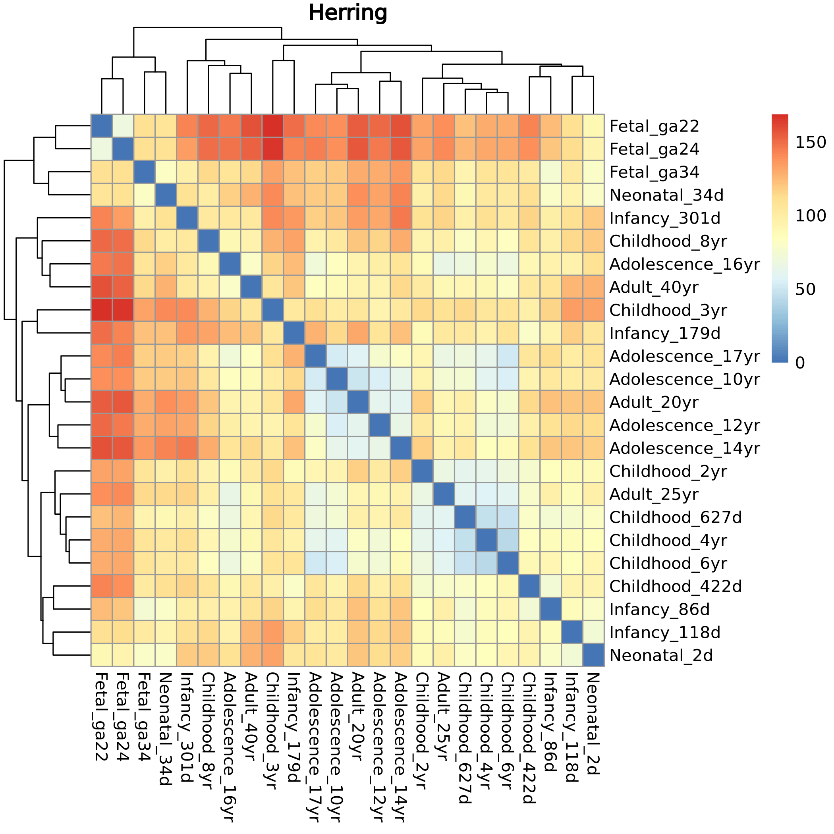
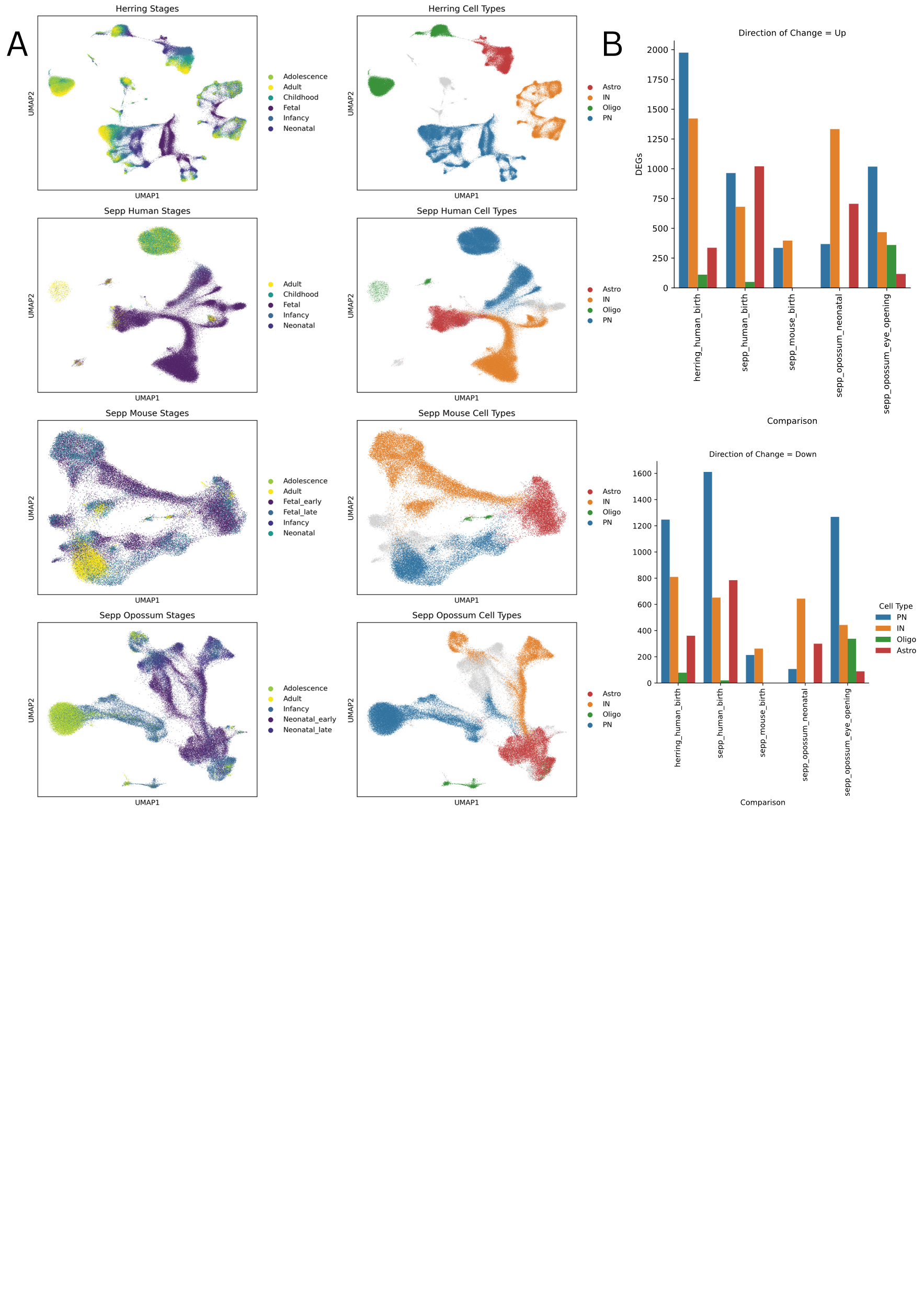


Figure 1: Heatmaps showing sample-to-sample distances for each dataset based on a variance stabilizing transformation applied to pseudobulks aggregated by sample age. Colouration shows the distance between samples, with greater values denoting more dissimilarity.

Another significant change was present in the mouse and opossum data between infancy and adolescence, coincident with eye-opening which occurs at approximately postnatal day 12 in mice (Tan *et al.* 2021) and day 34 in opossums (Macrini 2004; Djavadian *et al.* 2006). This suggested that the most significant transcriptomic changes occurring in our data were at the perinatal transition in human and mouse samples (and near birth for opossum), and eye-opening in mouse and opossum, consistent with prior studies. We used the clustering patterns from the sample-to-sample distance matrices to inform which samples we would include or exclude from each developmental group to best align them to our time points of interest, based on similarity with neighbouring groups. Samples which were part of the target birth or eye-opening developmental groups, or samples at the boundary of adjacent groups which were more similar to samples in the target groups, were excluded to better capture changes across those periods. Once these groupings had been set, cells were pseudobulked by sample age and cell type (“agewise” comparisons) and by developmental stage, sample age and cell type (“stagewise” comparisons). All datasets are visualised using Uniform Manifold Approximation Projection (UMAP) plots in Figure 2a, highlighting age progressions and cell type groupings.

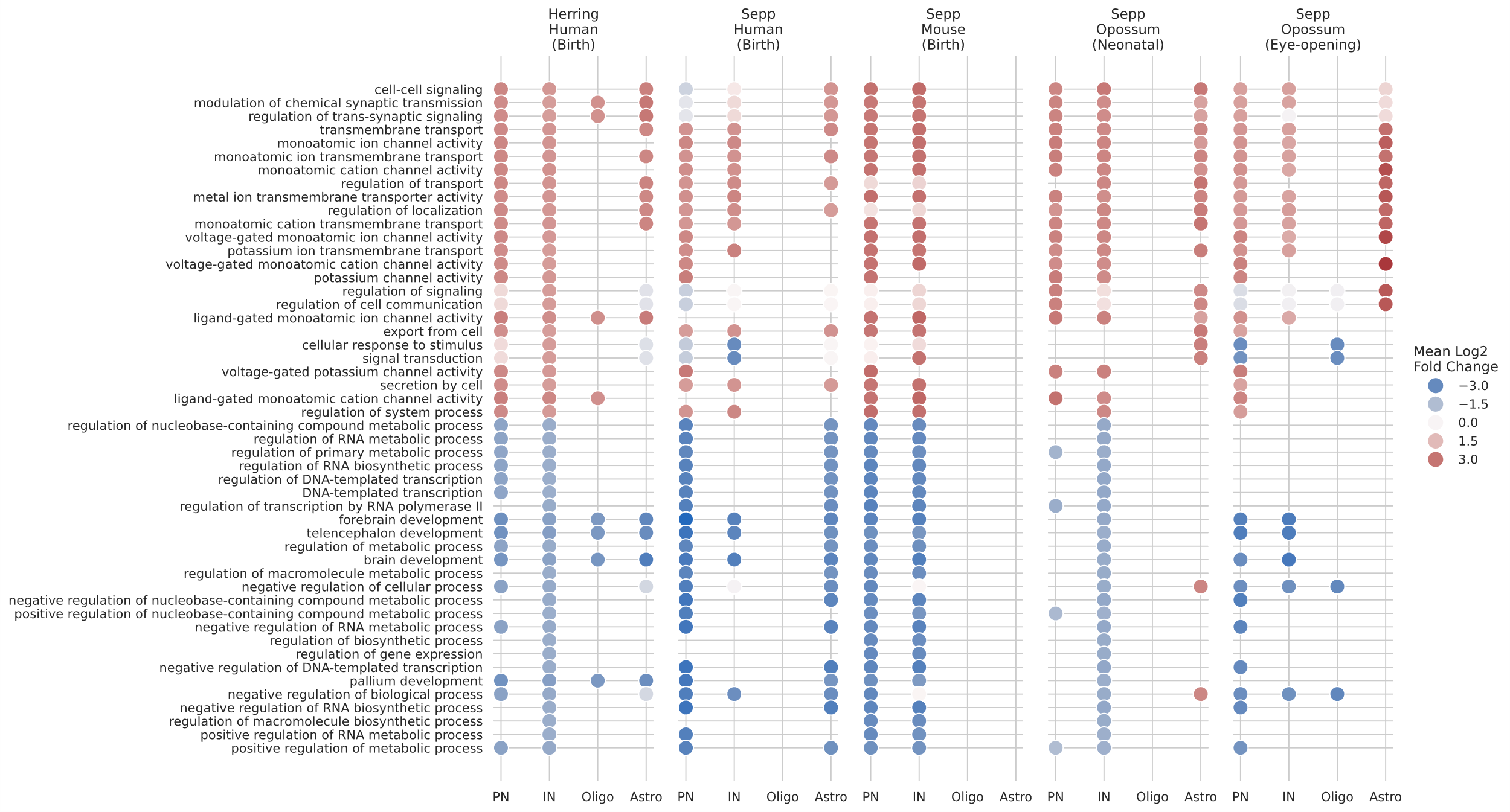
Differentially expressed genes (DEGs) were called using the standard DESeq2 pipeline (Love, Huber and Anders 2014) on each of these pseudobulks to generate results at two different developmental resolutions - age and broader developmental stage. While the stagewise approach more robustly captures differences across periods relevant to our hypothesis by considering multiple samples of similar developmental age and accounts for differences between cell types, both approaches were necessary to obtain results for some comparisons which were not possible with the stagewise approach alone. For agewise comparisons, all cell types were included in the analysis for each sample, so these results may be confounded by changes in cell type proportions between samples. In contrast, stagewise comparisons were made for principal (excitatory) neurons, inhibitory neurons, oligodendrocytes and astroglia separately to match nuclei based on broad function and capture possible lineage-specific differences in expression patterns. The Sepp datasets lacked oligodendrocytes and astrocytes at key developmental stages, making comparisons in these cell types difficult. Additionally, the vast majority of the astroglia group in the Sepp data consisted of radial glia (glial precursor cells) rather than true astrocytes as in the Herring data, meaning these groups did not match up well across datasets.

**Figure 2:** **a** UMAP embeddings for each dataset used in this study, coloured by developmental stage (left) and broad cell type (right). **b** Number of significant DEGs with a log2 fold change of at least 1 which were up- (top) or down- (bottom) regulated in each key comparison of the stagewise results.

For these reasons, we chose to restrict the majority of our analysis to just the neuronal cell types, which were far more abundant and more comparable across datasets. Limitations in the available data precluded a stagewise comparison of samples across eye-opening in mouse and across birth in opossum, as these both contained only a single sample in one of the comparison groups. The latter comparison was substituted for an early neonatal versus late neonatal comparison to keep as close to birth as the data would allow, but it should be noted that this does not properly capture a period equivalent to the birth comparison in the other species. The number of DEGs called for each key comparison from the stagewise approach are summarised in Figure 2b.

Gene Ontology Term Enrichment

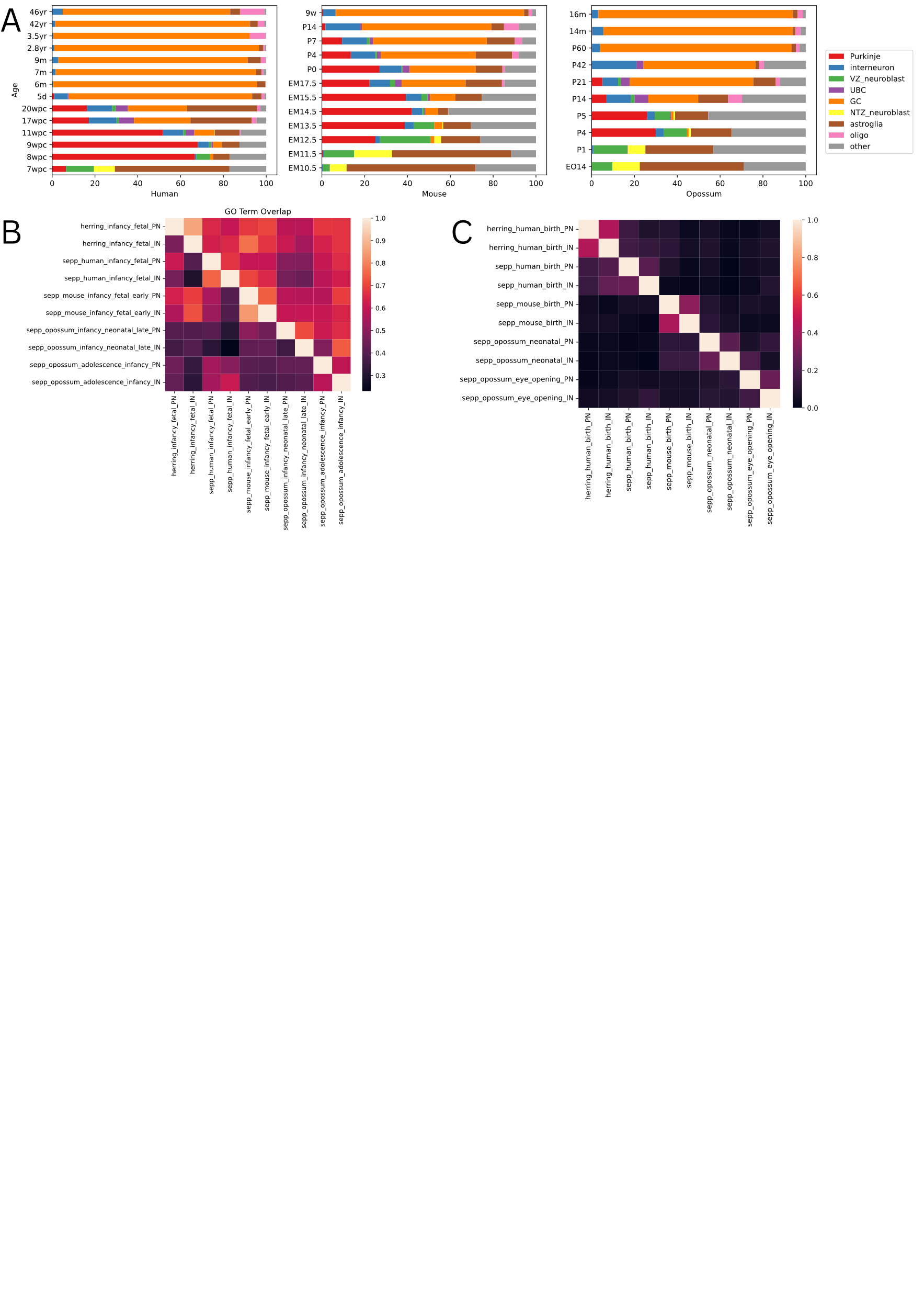
Having established when changes were occurring, we next sought to investigate what kinds of genes were changing at our time points of interest and determine if there were similarities across brain regions, cell types and species. Gene Ontology (GO; Ashburner *et al.* 2000; The Gene Ontology Consortium *et al.* 2023) term enrichment was performed using significant DEGs (adjusted p-value of <= 0.05) with a minimum log2 fold change of 1 via G:Profiler (Kolberg *et al.* 2023). Up- and down-regulated DEGs were queried separately for each comparison to elucidate expression changes in both directions. To provide a baseline for comparison between different groups, we took 50 of the most highly enriched terms from neuronal cell types in the Herring data across birth and calculated mean changes in expression for genes associated with those terms (Figure 3). In general the up-regulated terms are associated with cell signalling, cellular communication pathways and ion transport, while the down-regulated terms are associated with the regulation of various metabolic and biosynthetic processes, particularly those associated with RNA and DNA, cellular differentiation and tissue development. These patterns are loosely present across all species around birth, though concordance across datasets and cell types varies substantially. Neuronal cell types show much higher concordance than non-neuronal types, likely due to major differences in cell type proportions in each dataset (Figure 4a) and a dearth of non-neuronal nuclei in the Sepp data at relevant time points (Supplementary Figure 1), as well as reflecting functional differences between the different lineages. Examining the top 25 most up- and down-regulated terms for each comparison (Supplementary Table 1) confirms that despite the number of missing data points, which indicate a term was not significantly enriched in that group, similar terms with functions highly related to those seen in the Herring data at birth were enriched across all comparisons, supporting our primary hypothesis and suggesting that the changes in expression seen at birth are driven by a switch away from tissue development and growth towards setting up functional gene networks to prepare for neuronal activity. Overlaps between the top 150 GO terms for the neuronal cells in each dataset show much higher concordance in up-regulated terms than down-regulated terms, with no discernible consistent patterns in similarity between species (Figure 4b). The enrichment patterns seen at birth extend to the opossum at eye opening, however the down-regulated set at this time includes more terms related to cell cycle control. These findings were recapitulated in the agewise comparisons, where the mouse at eye-opening showed a similar tendency towards down-regulation of cell cycle control terms (Supplementary Table 2).

**Figure 3:** Dotplot of mean change in expression for genes associated with the 25 most enriched up- and down-regulated Gene Ontology terms in the Herring human dataset across birth. Only differentially expressed genes with an adjusted p-value of <= 0.05 and a log2 fold change of at least 1 are included. PN = principal neurons (excitatory neurons); IN = inhibitory neurons; Oligo = oligodendrocytes; Astro = astroglia.

Together, this suggests that the changes occurring around birth across all 3 species are similar, however at eye-opening the rodent species may have a further, unique switch away from regulating the cell cycle. In addition to matching genes to GO terms, G:Profiler also matches genes to transcription factor binding site motifs using TRANSFAC (Matys *et al.* 2003). Similar to GO term enrichment, this provides evidence for statistically enriched transcription factor binding site motifs among the queried genes. Comparing the enriched motifs for DEGs across the human and mouse results around birth revealed 9 commonly enriched motifs in the up-regulated DEGs and 57 in the down-regulated DEGs (Supplementary Table 3). No TRANSFAC results were returned for the opossum queries, preventing a comparison across all 3 species. Among the common motifs were three Krüppel-like factors (KLFs); GKLF, FKLF and LKLF; in the up-regulated queries and matches for several different Sp-family motifs in the down-regulated queries. KLFs and Sp-like factors are closely related families of zinc finger proteins which are widely expressed and have a variety of cell regulatory functions, including growth, proliferation, differentiation and developmental processes (Kaczynski, Cook and Urrutia 2003). KLFs have also been associated with a number of neuropsychiatric disorders, including Alzheimer’s, epilepsy and schizophrenia (Yin *et al.* 2015; Santos *et al.* 2024), suggesting an important role in proper brain development. Enrichment for several of the motifs for these transcription factors in significant DEGs across all three species provides evidence that they may have a role in regulating or driving the gene expression changes around birth.

Differentially Expressed Genes

In an effort to more objectively measure the similarity in expression between different species and stages and provide further evidence to support the GO term enrichment findings, we sought to examine the overlap in up- and down-regulated DEGs across each comparison. DEGs were filtered to include only the 7634 genes which were common to all datasets, then filtered further to include only those which were statistically significant (adjusted p-value of <= 0.05) with a log2 fold change of at least 1 and separated into up- and down-regulated sets for excitatory and inhibitory neurons for each comparison. This produced gene sets of differing sizes, so only the top 100 most significant genes in each set were considered to ensure proportions were comparable. Overlaps were then calculated as the proportion of genes common to each pair of gene sets and visualised using a heatmap (Figure 4c). These showed low overall concordance between different species and between different cell types in the same species, with a maximum overlap of 44% between down-regulated genes in the Herring excitatory and inhibitory neurons. Weak patterns of higher correlation can be seen between the two human datasets and between the two rodent datasets around birth, with minimal correlation across human and rodent datasets. This suggests that the genes most active in humans and rodents around birth are largely unique and there are substantial differences in the specific gene networks changing between the prefrontal cortex and cerebellum in the human brain and between excitatory and inhibitory neurons within each region, though the general functions of these networks may be similar, as evidenced by the GO term enrichment results.

**Figure 4:** **a** Proportions of major cell types in each of the Sepp datasets. **b** Heatmapofoverlaps as a proportion of the top 150 Gene Ontology terms in neuronal cells for each dataset. **c** Heatmap of overlaps as a proportion of the top 100 differentially expressed genes in neuronal cells for each dataset. For heatmaps the top half (above the middle diagonal) represents up-regulated genes, while the bottom half represents down-regulated genes.

To further investigate the possible regulatory mechanisms which may be driving changes around birth we took a more targeted approach and looked for conserved changes occurring in genes associated with three domains: histone modification/epigenetic regulation, as sweeping changes in the epigenome are associated with this period (Lister *et al.* 2013), oxygenation/response to oxygen species, due to the transition from the prenatal to postnatal environment, and transcription factors, which are the primary effectors of genetic change. Results for each domain, separated by time point, broad neuronal cell type and direction of change for each dataset are provided in Supplementary Tables 4, 5 and 6.

|  |  |  |  |
| --- | --- | --- | --- |
| Gene Symbol | Cell type | Direction of Change | Function |
| NAT8L | IN | Up | Enzyme involved in synthesis of N-acetylaspartate acid, a prominent CNS metabolite implicated in a wide range of neurological disorders |
| TRIM37 | IN | Up | Epigenetic repressor of positive regulators for centriole reduplication |
| BHLHE41 | IN | Up | Transcriptional repressor involved in regulating circadian rhythm and cell differentiation |
| RYR2 | PN | Up | Calcium channel required for calcium ion homeostasis |
| PTPRK | IN | Up | Signalling molecule which regulates cell growth, differentiation and mitotic cycle |
| GRIN1 | IN | Up | Subunit for a family of glutamate receptor channels involved in synaptic plasticity, linked to learning and memory |
| ADCY8 | IN | Up | Enzyme involved in cAMP pathway which regulates synaptic plasticity |
| MEIS2 | PN | Down | Transcriptional regulator, member of the TALE family which are important in developmental programs |
| RFX4 | IN | Down | Transcription factor involved in early brain development |

**Table 1:** **Summary of statistically significant differentially expressed common to humans and at least one of mouse and opossum perinatally.** PN = principal neurons (excitatory neurons); IN = inhibitory neurons. Functions are summarised from the relevant GeneCards entry; <https://www.genecards.org/> (Stelzer *et al.* 2016).

In the epigenetic results, a total of five up-regulated and 11 down-regulated genes were commonly differentially expressed between the human datasets. Of these, three were present in the opossum, all of which were up-regulated and found in the inhibitory neurons. None were also present in the mouse DEGs. Similarly, for oxygenation genes, nine up-regulated and six down-regulated DEGs were common to the human datasets, with three of these common across all species. For transcription factors, 19 up-regulated and 16 down-regulated DEGs were common in the human datasets, with another three shared between all species. DEGs common to more than one species are summarised in Table 1, representing the genes which are the best candidates as driving perinatal changes based on conserved patterns of expression across brain regions and species. The functions and direction of change for these genes also match well to the rest of the findings of this study, with the up-regulated genes being those associated with neuronal function or repressors of cell or tissue development, while the down-regulated genes are involved in development. The dearth of DEGs with common expression patterns across brain regions and species further points to substantial differences in the specific gene networks involved, consistent with prior analysis.

**Discussion**

This study aimed to characterize the gene expression changes occurring around birth in the human brain, determine whether these are consistent across brain regions and conserved across species, and present candidates for genes which may be driving these changes. Differential gene expression and GO term analyses suggest that while the general functions of the active gene networks around birth in all three species are similar, with a consistent switch away from tissue development and towards more specific functions associated with neuronal activity, the specific genes implicated share little similarity. Despite this, we found a limited number of DEGs with conserved expression patterns across species and provide some evidence for KLF and Sp-like protein involvement, which may be interesting targets for future studies, especially those focussed on advancing organoid maturation strategies. These findings are limited by several key factors. DEG comparisons were limited by the number of genes common to all datasets, which was largely constrained by the opossum data. Although there were over 20000 gene features in the opossum data, approximately 11500 of these were putative genes with an XLOC identifier (see Methods). As such these have no known orthologues and could not be used in comparisons with the other species. Similar studies may consider replacing the opossum dataset with that of a different species with a more completely annotated genome, which would expand the common gene space and allow a greater number of genes to be included in comparisons. Differential gene expression results were also limited by the design of the datasets, which all contained several potential confounding factors, such as sample sex, race, nuclei preparation method, library prep lot number, etc., which could not be properly controlled for as there was only a single replicate at each sample age. This made it impossible to untangle these effects from those due to age, which could be achieved in future datasets by making sure there was a replicate for each category of all major cofactors present at every sample age, though the difficulties inherent in acquiring samples for complex datasets of this nature may make that infeasible. Similarly, differences in the age distributions of samples in each dataset made it difficult to meaningfully compare birth and eye-opening in the rodent samples, since the opossum data had only a single prenatal sample which was extremely close to birth, and the mouse data had a single sample post eye-opening which skipped a large portion of development. This prevented these two key periods from being included at a single developmental resolution, stifling attempts to properly compare them in ways which would reveal whether the changes seen at each stage were more or less similar to those at birth in humans. Additionally, biases in the cell type proportions at key stages between datasets reduced our ability to make meaningful comparisons for some cell types, hence our focus on the neuronal cell types in the analysis. This was especially prevalent in the oligodendrocyte populations for the Sepp data, which were depleted or entirely absent at some developmental stages. This also affected the astroglia in the Sepp data, where the majority of the nuclei appeared to be radial glia rather than true astrocytes, based on marker gene expression (see Methods). Bespoke datasets built with the unique demands of comparative analysis in mind, such as better balancing the relevant confounding factors, would greatly benefit similar analyses and allow a more robust characterization of the changes occurring at these crucial time periods. Since this study is purely bioinformatics based, further work is required to confirm the effects of the candidate genes presented with in vivo studies and functional assays.

**Methods**

For this study the Herring data was used as a benchmark due to the thoroughness and transparency of its processing documentation and overall higher quality by typical quality control metrics. The preprocessing procedures were focused on bringing the Sepp data to parity using methods largely based on the Herring processing procedures, with adjustments made to manage the unique demands of the dataset, to ensure the datasets matched as best as possible and facilitate meaningful comparisons between them. All code was run inside a conda environment (conda contributors 2025), which can be reproduced using the provided .yml file (see Data Availability).

Data Pre-processing

The processed Herring data was downloaded directly from a Google bucket linked on the Lister Lab website (brain.listerlab.org; see Data Availability) as an anndata (Virshup *et al.* 2024) object which includes all metadata, raw counts, downsampled and normalized counts, and a variety of miscellaneous supplemental data for human pre-frontal cortex samples. Raw Sepp data was downloaded from the Kaesmann Lab repository (see Data Availability) as RDS objects containing only the raw counts for human, mouse and opossum cerebellum samples in SingleCellExperiment format (Amezquita *et al.* 2020). All genes in the Sepp data were identified using their Ensembl ID or an arbitrary XLOC identifier present only in the opossum data. Ensembl IDs were converted to their appropriate gene symbol using the BiomaRt R package (Durinck *et al.* 2009), stored alongside their Ensembl ID and the resulting matrix was converted to anndata using the anndata R package (Wolf, Angerer and Theis 2018) . XLOC identifiers were confirmed by Dr Mari Sepp (personal communication – Sepp 2024) to derive from another study (Wang *et al.* 2020) and represent putative transcripts which do not map to known genes and therefore do not have corresponding gene symbols. A pre-mRNA gene transfer format file was provided, containing the genomic positions for these putative genes (see Data Availability). Metadata contained in Supplementary Table 1 of the Sepp paper was adjusted to convert the sample ages given in Carnegie stages to weeks post conception based on conversions given in the body of the paper. These were converted again to a common numerical scale consistent with the Herring data and all 3 designations were included in the final metadata for completeness. All resulting metadata was added to the count matrix to build the complete anndata objects for the Sepp data.

Cleaning and Filtering

Since the processing that had already been performed on the Sepp data was unclear, steps were taken to check key features and ensure the quality standard was as consistent as possible with the Herring data. Normalization did not appear to have been applied based on the available documentation and sums of the raw count data producing integer results. Common normalization processes based on log transformations or regression models introduce floating point values into the count matrix, making this a quick way of determining whether they have been applied. Genes present in less than 5 cells were removed as uninformative, then feature counts per cell and library sizes were plotted for each batch to visualise their distributions and identify any low-quality runs. Nuclei were screened based on feature count and any with counts lower than 3 median absolute deviations below the mean were discarded. Cells were screened for mitochondrial and ribosomal gene content and found to contain much higher proportions of ribosomal RNA than mitochondrial. Since both represent ambient RNA contamination they are expected to be correlated, however the very low overall mitochondrial gene content suggested that prior filtering had already been applied to remove mitochondrial RNA, so no extra filtering was performed on this basis. Cells with ribosomal gene counts greater than 3 median absolute deviations above the mean, as a percentage of total cell counts, or with greater than 20% total ribosomal gene content were discarded. Genes were screened based on expression as a percentage of total counts and any that were deemed problematically dominating were removed, including MALAT1 and its orthologues which are known to be highly captured in poly-A single cell RNA assays (Clarke and Bader 2024). Doublet detection had been already performed for all datasets according to their respective documentation and was not run again to prevent removal of false positives.

Normalization

Downsampling was applied to the Sepp data to mitigate bias due to library size. Any cells with total counts less than 1000 were removed and those remaining were downsampled to 1000 counts per cell. Two separate transformations were applied to the downsampled counts and stored concurrently. First, counts were binarized to represent whether each feature was present or not in each cell, for use in cell typing where we are more interested in whether marker genes are expressed or not, rather than the magnitude of their expression. Second, a log1p transformation was applied to counts to reduce the range of expression values while preserving variance, intended for use in all other analytical applications which did not require the use of raw count data.

Dimensionality Reduction and Feature Selection

Feature selection was performed using the Scanpy (Wolf, Angerer and Theis 2018) implementation of the Seurat v3 algorithm for identifying the most highly variable genes in each of the Sepp datasets, selecting for the top 5000 most deviant genes. This set of genes represents the most informative features of each dataset and was used for the remainder of the preprocessing steps to make the data more tractable for visualisation. Each dataset was then put through a series of Scanpy functions using the downsampled, binarized counts for the highly variable genes to embed the data into a lower dimensional space and better visualise broad patterns in expression. Principal components were calculated using the Principal Component Analysis (PCA) function, which were then used to calculate a nearest neighbour graph as the basis for a UMAP embedding (Supplementary Figure 2). Colourations were applied to the UMAP embedding based on different aspects of metadata, including sample tissue, sex, nuclei preparation method and 10x protocol version, to visualise possible batch effects that may negatively impact the results. In the mouse data some batches consisted of pooled samples from more than one individual. Batches with pooled male and female samples did not integrate on the UMAP embedding and were removed (Supplementary Figure 3).

Cell Type Annotation and Trajectory Verification

The Sepp data already contained cell type labels at several different levels of precision based on typical clustering and annotation strategies, coupled with some extra information from accompanying spatial data, e.g. FISH, and morphological data, mostly used for subtyping of cells expressing non-canonical markers (Sepp *et al.* 2024). To verify these annotations we used known marker genes taken from existing literature (Supplementary Table 7) to make sure expression patterns generally conformed to what we expected to see, identify any potential further sources of bias, and ensure we could be confident in the labels assigned. To this end, we determined that the highly specific subtypes which were already annotated were unnecessarily granular for the analysis we sought to achieve, so more general labels representing the major cerebellar cell types were verified instead. UMAPs from the previous step were integrated using Harmony (Korsunsky *et al.* 2019; Supplementary Figure 2) to clarify cell type groupings and trajectories, then visualised to compare the existing broad cell type labels with the expression of relevant marker genes (Supplementary Figures 4, 5, 6).

Differential Gene Expression Analysis

Agewise comparisons were performed by pseudobulking each dataset by sample age and broad cell type which was then filtered to select data corresponding to each pairwise combination of sample ages in the dataset. Each combination was converted to a DESeqDataSet object using a design variable including both sample age and cell type, then run through the standard DESeq2 analysis pipeline to produce a results dataframe. All cell types were included in each call to the DESeq2 pipeline, as separating them failed to produce results due to the resulting matrix being computationally singular. This was a limitation of the experimental design in our datasets and caused by a lack of replicates at each age. Similarly, Surrogate Variable Analysis (Leek and Storey 2007) and ComBat (Johnson, Li and Rabinovic 2007), methods recommended for covariate control in the DESeq2 documentation, both failed due to computationally singular matrices.

For stagewise comparisons, datasets were pseudobulked by sample age and developmental stage, converted to DESeqDataSet objects and transformed using a variance stabilizing transformation, after which distance values were calculated to determine the degree of difference between samples at each age, as set out in the DESeq2 documentation. These matrices were visualised and used to help determine the appropriate stage groupings for each dataset (Figure 1), as a measure of the change in expression occurring between samples. On this basis, stage labels were adjusted to best reflect the expression patterns seen and key comparison groups were selected to ensure we were capturing the fullest extent of the changes occurring, especially at the perinatal transition. Following this, a similar procedure to the agewise comparisons was used, however data from each broad cell type was separated out and used as input to a separate call to the DESeq2 pipeline.

For the analysis of these results, overlaps in the DEGs of key comparisons were defined as the proportion of significant DEGs common to the up- and down-regulated groups for principal and inhibitory neurons across datasets. These were limited to an equal number of DEGs with the lowest p-values in each group, to ensure gene sets of the same size and enable proportions to be meaningfully compared. Searches for genes associated with particular functional domains were conducted by matching search terms to the Gene Set Enrichment Analysis ontology gene sets for humans and mice (version 2024.1; <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>), where any genes that were members of the returned gene sets were used.

Gene Ontology Term Enrichment Analysis

Each DEG dataframe for both agewise and stagewise comparisons was filtered to include only DEGs with an adjusted p-value of <= 0.05 and a log2 fold change of at least 1, to ensure only statistically significant results were used and remove lowly expressed genes with changes that were not biologically meaningful. These were used as separate inputs to G:Profiler via the Scanpy wrapper to produce a matching dataframe with GO term enrichment results for each DEG dataframe. The top 25 most up- and down-regulated terms across the neuronal cell types in the Herring data at birth were combined into a single list of 50 terms, which was used as the benchmark for comparisons with the other datasets at birth and eye-opening. Many of the most enriched terms were quite broad, consisting of varied biological processes with large gene sets which were naturally more likely to intersect with a greater number of DEGs and thereby be called as highly enriched. Some extremely broad terms, such as ‘multicellular organismal process’, appeared as highly enriched in both the up- and down-regulated groups, making them uninformative. To combat this, terms were filtered to include only those with a minimum of two parent terms, ensuring a minimum level of specificity to prevent them being too broad, and those which were only present in either the up- or down-regulated groups, but not both, as these provided the clearest idea of the differences between the groups. Mean changes in expression for each term were calculated by scraping the intersecting DEGs for that term from the relevant GO term enrichment result dataframe if it was present, then averaging the log2 fold change for all intersecting DEGs based on the values in the corresponding DEG dataframe. Overlaps in enriched GO terms were calculated in the same way as overlaps in DEGs, described above.

**Data Availability**

All code used for running the preprocessing, quality control and analysis for this paper, a .yml file for the conda environment used, along with Supplemental Figures and Tables are provided at the GitHub repository https://github.com/cmcphan/SCIE5574\_thesis.

Processed Herring data available at https://console.cloud.google.com/storage/browser/neuro-dev/Processed\_data;tab=objects?invt=AbuMnQ&prefix=&forceOnObjectsSortingFiltering=false

Raw Sepp data available at <https://apps.kaessmannlab.org/sc-cerebellum-transcriptome/> under the ‘Info, Methodology & Data’ tab. Pre-mRNA versions for each species were used, listed under the ‘Data & code access’ section.

**Acknowledgments**

Thank you to my supervisors Chuck Herring and Ryan Lister for facilitating this work. An extra thank you to Chuck Herring who helped support and guide all aspects of this project from its inception all the way through to writing. Thank you to Rohini Trivedi for assisting with proofreading of this manuscript.

**References**

Amezquita RA, Lun ATL, Becht E *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* 2020;**17**:137–45.

Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:1–12.

Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Bakken TE, Miller JA, Ding S-L *et al.* A comprehensive transcriptional map of primate brain development. *Nature* 2016;**535**:367–75.

Bendl J, Fullard JF, Girdhar K *et al.* Chromatin accessibility provides a window into the genetic etiology of human brain disease. *Trends Genet* 2025, DOI: 10.1016/j.tig.2025.01.001.

Camp JG, Badsha F, Florio M *et al.* Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci* 2015;**112**:15672–7.

Clarke ZA, Bader GD. MALAT1 expression indicates cell quality in single-cell RNA sequencing data. 2024:2024.07.14.603469.

conda contributors. conda: A system-level, binary package and environment manager running on all major operating systems and platforms. 2025.

Czarny P, Bialek K, Ziolkowska S *et al.* DNA damage and repair in neuropsychiatric disorders. What do we know and what are the future perspectives? *Mutagenesis* 2020;**35**:79–106.

Djavadian R, Bisti S, Maccarone R *et al.* Development and plasticity of the retina in the opossum Monodelphis domestica. *Acta Neurobiol Exp (Warsz)* 2006;**66**:179–88.

Durinck S, Spellman PT, Birney E *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91.

Herring CA, Simmons RK, Freytag S *et al.* Human prefrontal cortex gene regulatory dynamics from gestation to adulthood at single-cell resolution. *Cell* 2022;**185**:4428-4447.e28.

Huang S-Y, Ge Y-J, Ren P *et al.* Genome-wide association study unravels mechanisms of brain glymphatic activity. *Nat Commun* 2025;**16**:626.

Jeong E, Choi S, Cho S-W. Recent Advances in Brain Organoid Technology for Human Brain Research. *ACS Appl Mater Interfaces* 2023;**15**:200–19.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.

Joshi R, Salton SRJ. Neurotrophin Crosstalk in the Etiology and Treatment of Neuropsychiatric and Neurodegenerative Disease. *Front Mol Neurosci* 2022;**15**, DOI: 10.3389/fnmol.2022.932497.

Kaczynski J, Cook T, Urrutia R. Sp1- and Krüppel-like transcription factors. *Genome Biol* 2003;**4**:206.

Kim S, Chang M-Y. Application of Human Brain Organoids—Opportunities and Challenges in Modeling Human Brain Development and Neurodevelopmental Diseases. *Int J Mol Sci* 2023;**24**:12528.

Kolberg L, Raudvere U, Kuzmin I *et al.* g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res* 2023;**51**:W207–12.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;**16**:1289–96.

Leek JT, Storey JD. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet* 2007;**3**:e161.

Li M, Santpere G, Kawasawa YI *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 2018;**362**:eaat7615.

Li M-L, Tang H, Shao Y *et al.* Evolution and transition of expression trajectory during human brain development. *BMC Evol Biol* 2020;**20**:1–16.

Lister R, Mukamel EA, Nery JR *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* 2013;**341**:1237905.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21.

Luo C, Lancaster MA, Castanon R *et al.* Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep* 2016;**17**:3369–84.

Macrini TE. Monodelphis domestica. *Mamm Species* 2004:1–8.

Matys V, Fricke E, Geffers R *et al.* TRANSFAC ® : transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**:374–8.

Mulder LA, Depla JA, Sridhar A *et al.* A beginner’s guide on the use of brain organoids for neuroscientists: a systematic review. *Stem Cell Res Ther* 2023;**14**:1–21.

Repetto L, Chen J, Yang Z *et al.* The genetic landscape of neuro-related proteins in human plasma. *Nat Hum Behav* 2024;**8**:2222–34.

Santos AB, Carona A, Ettcheto M *et al.* Krüppel-like factors: potential roles in blood-brain barrier dysfunction and epileptogenesis. *Acta Pharmacol Sin* 2024;**45**:1765–76.

Sepp M. Cellular Development Dataset Clarification. 2024.

Sepp M, Leiss K, Murat F *et al.* Cellular development and evolution of the mammalian cerebellum. *Nature* 2024;**625**:788–96.

Stelzer G, Rosen N, Plaschkes I *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinforma* 2016;**54**:1.30.1-1.30.33.

Tan L, Ma W, Wu H *et al.* Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* 2021;**184**:741-758.e17.

The Gene Ontology Consortium, Aleksander SA, Balhoff J *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* 2023;**224**:iyad031.

Virshup I, Rybakov S, Theis FJ *et al.* anndata: Access and store annotated data matrices. *J Open Source Softw* 2024;**9**:4371.

Wang Z-Y, Leushkin E, Liechti A *et al.* Transcriptome and translatome co-evolution in mammals. *Nature* 2020;**588**:642–7.

Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15.

Yin K-J, Hamblin M, Fan Y *et al.* Krüpple-like factors in the central nervous system: novel mediators in Stroke. *Metab Brain Dis* 2015;**30**:401–10.

Zhu Y, Sousa AMM, Gao T *et al.* Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* 2018;**362**:eaat8077.