Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by:

M.Sc. Kevin Leiß

born in: Seeheim-Jugenheim, Germany

Oral examination: 22.03.2023

Development and Evolution of the Mammalian Cerebellum at Single Cell Resolution

Referees: **Prof. Dr. Henrik Kaessmann**

Prof. Dr. Dr. Georg Stoecklin

Abstract

Originally thought to only take part in motor control, the cerebellum emerged over the last decades as an important organ in various higher cognitive functions, such as learning and speech [1]. Besides this, the cerebellum is associated to various diseases, such as spinocerebellar ataxia, autism spectrum disorder, and medulloblastoma [2]. The basic structure and connective properties of it are well understood, but single-cell-technologies made it possible to study the cerebellum at higher resolution. Many questions about molecular details of its development and evolution are still not answered. Cerebella are present in all jawed vertebrates, though structural diversity is macroscopic and microscopic detectable, such as the number of deep nuclei, the presence of the vermis, or the mode of production of one of the most important cell types in the cerebellum - granule cells [2–4].

Using single-nucleus RNA-sequencing (snRNA-seq) and bioinformatic approaches, I studied cerebellum data of human, mouse (*Mus musculus*) and opossum (*Monodelphis domestica*). The dataset contained samples spanning the organs development at high temporal resolution. It was possible to track the differentiation of the major cerebellar neuronal and glial cell types, as well as identify states and subtypes. This generated a comprehensive map of cellular complexity through eutherian (human and mouse) and marsupial (opossum) development. Leveraging the evolutionary distance of approximately 160 million years between the eutherian and marsupial lineage, conserved and diverged cell type marker genes could be identified which might be promising candidates for understanding the basic blueprint of cerebellar cell type identity.

Stage correspondence mapping aligned the vastly different developmental time frames of the three studied species and allowed the identification of a two-fold increase in Purkinje cell progenitors in the human lineage, which might be connected to a recently identified human-specific secondary ventricular zone progenitor pool [5].

It was possible to model the differentiation path of granule and Purkinje cells from early progenitors to mature neurons. Conserved and diverged gene expression trajectories were discovered. Using *in vitro* and *in vivo* intollerance scores [6], I could show that genes which are dynamically expressed during differentiation show higher functional constraint as non-dynamic genes, fitting to previous bulk-RNA-seq studies [7], showing similar results across the development of the full organ. Some orthologs with diverging patterns were disease-associated genes, which could have implications on clinical research on conditions like autism spectrum disorders and medulloblastoma.

Furthermore, fundamental changes of gene expressions, established as gain or loss of expression within a cell type and species, were detected. Affected genes showed decreased functional constraint, verifying evolutionary principles on single cell scale [8].

Taken together, this study shows the strength of state of the art methodology combined with high resolution developmental sampling in an evolution biological context to discover fundamental

principles of organ development at single-cell scale.

Zusammenfassung

Ursprünglich dachte man, dass das Kleinhirn (Cerebellum) nur an motorischer Kontrolle beteiligt ist, aber in den letzten Jahrzenten erkannte man, dass es auch bei der Verarbeitung höherer kognitiver Funktionen wie z.B. Lernen und Sprache beteiligt ist [1].

Außerdem steht das Kleinhirn in Verbindung mit verschiedenen Krankheiten wie der spinozerebellären Ataxie, der Autismus-Spektrum-Störung und dem Medulloblastom, einem pädiatrischen Hirntumor [2]. Die Grundstruktur und die Verbindungseigenschaften des Kleinhirns sind gut verstanden, aber die Einzelzelltechnologien ermöglichten, das Kleinhirn detaillierter zu untersuchen. Viele Fragen zu molekularen Details seiner Entwicklung und Evolution sind noch nicht vollständig beantwortet. Kleinhirne sind bei allen Gnathostomata vorhanden, doch strukturelle Unterschiede sind makroskopisch und mikroskopisch nachweisbar, wie etwa die Anzahl der sogenannten "deep nuclei", das Vorhandensein der Vermis oder die Produktionsweise eines der wichtigsten Zelltypen im Kleinhirn - den Körnerzellen [2–4].

Mithilfe der Einzelkern-RNA-Sequenzierung (single-nucleus RNA-seq, snRNA-seq) und bioinformatischer Ansätze habe ich Kleinhirndaten von Mensch, Maus (*Mus musculus*) und Opossum (*Monodelphis domestica*) untersucht. Der Datensatz enthielt Proben, die die Entwicklung des Organs in hohe zeitliche Auflösung abbilden. Es war möglich, die Differenzierung der wichtigsten neuronalen und glialen Zelltypen des Kleinhirns zu verfolgen, sowie Zellstadien und Subtypen zu identifizieren. So entstand eine umfassende Karte der zellulären Komplexität des Kleinhirns während der Entwicklung von Eutheriern (Mensch und Maus) und Beuteltieren (Opossum). Unter Ausnutzung des evolutionären Abstands von etwa 160 Millionen Jahren zwischen der Eutherier- und Beuteltierlinie konnten konservierte und divergierende Zelltyp-Markergene identifiziert werden, die vielversprechende Kandidaten für das Verständnis des grundlegenden Bauplans der Kleinhirn-Zelltypen sind.

Durch Verfahren zur Angleichung der sehr unterschiedlichen Entwicklungszeiträume zwischen den drei untersuchten Spezies, konnte eine doppelte Abundanz an sich entwickelnden Purkinje Zellen im Menschen detektiert werden. Es wäre möglich, dass diese Beobachtung im Zusammenhang mit einem jüngst identifizierten Vorläuferpool an der ventrikulären Zone steht [5].

Es konnte der Differenzierungsweg von Körner- und Purkinje-Zellen von frühen Vorläuferzellen zu reifen Neuronen modelliert werden. Es wurden konservierte und divergierende Genexpressionstrajektorien entdeckt. Mit Hilfe von *in vitro* und *in vivo* Intolleranzwerten konnte ich zeigen, dass Gene, die während der Differenzierung dynamisch exprimiert werden höheren funktionelle Einschränkungen aufweisen als nicht-dynamische Gene. Dies passt zu früheren RNA-seq-Studien [7], die ähnliche Ergebnisse bei der Entwicklung des gesamten Organs zeigten. Einige Orthologe mit divergierenden Mustern waren krankheitsassoziierte Gene, was Auswirkungen auf die klinische

Forschung zu Krankheiten wie Autismus-Spektrum-Störungen und Medulloblastom haben könnte.

Darüber hinaus wurden grundlegende Veränderungen der Genexpression, die als Zunahme oder Verlust der Expression innerhalb eines Zelltyps und einer Spezies definiert wurden, festgestellt. Betroffene Gene wiesen eine geringere funktionelle Einschränkung auf, was evolutionäre Prinzipien auf der Ebene der einzelnen Zelle bestätigen [8].

Alles in allem, zeigt diese Studie die Stärke einer modernen Methodik in Kombination mit temporär hochauflösenden Probenschemata in einem evolutionsbiologischen Kontext, um grundlegende Prinzipien der Entwicklung eines Organs auf Einzelzellebene zu erforschen.

Contents

1 Preface

The research presented in this thesis was only possible due to the combined effort of a highly motivated group of people. Please find the list of everybody involved in the preprint [10] and all supporting people in the Acknowledgement.

I would like to highlight Dr. Mari Sepp, who did all the heavy lifting, supported by the great team of the Kaessmann lab, in the wetlab, organized samples and helped in designing and interpreting the results you are about to see. Therefore, whenever I mention that "we" analysed or interpreted something in the following thesis, I express that Dr. Mari Sepp and I discussed the result and came to the presented result or interpretation. Modern research and especially big projects, like this one, are never the result of the work of one person but always rely on teamwork within and across labs around the world.

2 Introduction

Humans have been studying their own anatomy probably since the beginning of their existence. Written historic evidence for anatomical studies in surgical context is available from as early as 1,600 years BC and known as the Edwin Smith papyrus [11]. Helenistic Alexandria was the origin of ancient anatomical studies with anatomists and philosophers such as Aristotle, Praxagoras and Herophilus. As early as the second century BC, scholars invesitaged human major organs, among which the brain and its adjacent structures were discussed, even though their functions remained to be elucidated [12]. Posterior to the "brain", a smaller structure was observed - an organ which we know today as the cerebellum (diminutive form of the latin word *cerebrum*, which means brain, therefore the "small brain"). 400 years later, the cerebellum and its possible function was described by Claudius Galenus, also known as Galen, in the Roman empire. Galen suggested that the cerebellum might play a role in motor control by being the source of motor nerves and the spinal cord ([13], p. 629). Amazingly, this suggestion regarding the function of the cerebellum is not far from the current understanding of its main purposes.

2.1 Cerebellar functions and structure

In the modern age, the cerebellum has been mainly recognized to function in motor control and planning [14]. As early as the beginning of the nineteenth century, two scientists, named Pierre Flourens [15] and Luigi Rolando [16], observed that after lesions in the cerebellum, motor control was impaired in different areas of the body. They did not find obvious changes in intellectual capacity of their subjects, which made them believe that the cerebellum does not influence higher cognitive functions. In fact, this observation, even though, at the same time, the cerebellum is highly connected to cerebral association networks, is still a mystery to be resolved [1]. Mounting evidence suggest that the cerebellum is involved in memory, language and processing of sensory information (reviewed in [1, 17]). The role of the cerebellum in these non-motor control activities is also reflected in its association to diseases such as autism spectrum disorder and schizophrenia [4, 17], besides diseases affecting motor control such as spinocerebellar ataxia [4]. Pediatric brain tumor research is also focusing on the cerebellum, due to the development of embryonic and childhood cancers in this brain region. The cerebellum to the most often observed site of central nervous system cancers, which are the leading cause of cancer related deaths in children [18, 19]. The cerebellar tumors are divided to different types: medulloblastoma, ependymoma, and pilocytic astrocytoma [19].

The main macroscopic regions of the mammalian cerebellum are the two hemispheres and the vermis, located in between the hemispheres. The boundary between vermis and either of the hemispheres is called paravermis. Laterally, the cerebellar hemispheres extend to the paraflocculi and flocculi. Ten conserved lobules are visible, organized on the anterior - posterior axis. Histological

cuts through the organ reveal the foliation pattern of the adult cerebellum. The cerebellar cortex presents itself as a three-layered structure, which was described by the founding father of modern neurobiology, Ramon y Cajal, more than 100 years ago [20]. The outer layer, the so called molecular layer (ML), contains axons of granule cells, also called parallel fibers, Purkinje cell dendrites, interneurons (stellate and basket cells), and the termini of climbing fibers. Below the molecular layer, the Purkinje layer (PL) resides, which is only one cell layer thick. The namesake of this layer are the there located Purkinje cells. The innermost layer of the cerebellar cortex is called granule cell layer. Again, the name-giving granule cells are located in this area. Granule cells are the most numerous cell type of the brain (up to 80% in mammals) and have a comparatively small cell body [21]. Below the cerebellar cortex, buried inside the cerebellar white matter, the cerebellar nuclei are located.

The cerebellum has a highly stereotyped microcircuitry. Neurons of the inferior olivary complex of the brainstem project to Purkinje cells. Via granule cells, Purkinje cells also receive signals from mossy fibers which have their origin in nuclei in the brainstem and spinal cord. The connection between granule and Purkinje cells is mediated by bifurcated axons (parallel fibers in the molecular layer) which can be contacted by up to 300 Purkinje cells. Firing of Purkinje cells can be modulated by interneurons, such as stellate and basket cells [22]. Due to the high number of granule and very low number of Purkinje cells, incoming signals undergo extensive integration. Besides granule cells, unipolar brush cells (UBC) are modulating signals from mossy fibers to Purkinje cells. The processed signal is projected from Purkinje cells to the cerebellar nuclei, which in turn contact various parts of the cerebrum [21]. Buckner *et al.*, building on previously conducted work on cerebellar - cerebrum connectivity (reviewed in [24]), traced cerebellar connection to association networks and could identify regions of the cerebellum mapping to different parts of the body [23]. Interestingly, visual and auditory cortex were not found to be represented in the human cerebellum [23].

This comparably simple cytoarchitecture, foliation and connectivity circuits form only the first level of cerebellar complexity. In situ hybridization, trascriptomic and functional studies revealed a more complex architecture: along the parasagittal axis an antigen called zebrinII was identified in Purkinje cells, revealing a striped pattern [25]. Later studies identified zebrinII as aldolase C (ALDOC), which among other genes is differentially expresed within the detected stripes [21, 26] (White & Sillitoe figure 4 [21]). ALDOC-positive and negative regions are present in all mammalian and avian species [27]. The ALDOC stripes are not equally distributed but subcoordinated within four groups of lobules [21]. This highlights that the cerebellum shows intricate substructures besides the visible layering.

2.2 Development

The blueprint for cerebellar development is mainly based on mouse studies, therefore the following summary refers to the mouse developmental timeline. Early cerebellar primordia are established at around embryonic day 8.5 (E8.5). Two antagonizing homeobox transcription factors establish the midbrain / hindbrain boundary: Otx2, expressed in the midbrain, and Gbx2, expressed in the hindbrain. Where their effects cancel each other, the isthmic organizer is established. Fgf8 is the key organizing signalling molecule [28–30]. The timing of Fgf8 expression is tightly regulated to accomodate correct initiation and development of the cerebellar anlage [31]. Many transcription factors control the initiation, maintenance and development of the cerebellum, e.g., *Pax2*, *Pax5* [32], *En1*, *En2* [33].

To generate the different cell types, resident in the adult cerebellum, two progenitor zones are established: inhibitory GABAergic cells are descendants of the ventricular zone, excitatory glutamatergic neurons are born at the rhombic lip [34]. The progenitors of both pools are thought to be multipotent radial glia cells [35]. These progenitor zones are defined by two major transcription factors: pancreas-specific transcription factor 1a, Ptf1a, is expressed at the ventricular zone [36]. Atonal homolog 1, Atoh1, is active in cells of the rhombic lip [37, 38]. The loss of either of these genes results in the lack of GABAergic, or glutamatergic neurons, respectively [38–40].

Neurognesis at both progenitor zones follows a sequential program giving rise to the different cell types found in the adult cerebellum. The earliest cell types, born at the rhombic lip and ventricular zone at E10.5 to E11.5, are cerebellar nuclei neurons (GABAergic and glutamatergic). Shortly thereafter Purkinje cells are born and leave cell cycle latest at E13.5. They migrate from the ventricular zone along the processed of radial glia cells into clusters dispersed between the migrating cerebellar nuclei cells [21]. The decision on which cluster a Purkinje cell belongs to is influenced by birth date [41, 42]. The migration of Purkinje cells is modulated by Reelin [43]. Interea, Purkinje axon development and maturation occurs and can be detected as early as E14.5 [44, 45]. The establishment of the previously mentioned ALDOC-positive and -negative patterning is not yet fully understood, but some key factors have been reported: Ebf2, *En1*, and *En2*. It was shown that Ebf2 can supress phenotype [46, 47].

At the same time, the external granule cell layer (EGL) starts to form from cells originating at the rhombic lip, and engulfs the cerebellum at E15.5. The EGL is a transient structure in which granule cell progenitors proliferate, and is the exclusive source of granule cells [34, 37, 38]. Postmitotic differentiating granule cells migrate along radial glia cells through the Purkinje layer to the granule cell layer [48]. An interesting interplay between Purkinje cells and granule cells orchestrates the exit from cell cycle of granule cells via the secretion of SHH by Purkinje cells [21]. The granule cell neurogenesis can be traced up to postnatal day 14 (P14) and migration of

postmitotic granule cells is completed at P20 [21]. This protracted development of the cerebellum is thought to make it especially susceptible to be the origin of various neurodevelopmental disorders [49]. Unipolar brush cells (UBC) can be detected as early as E14.5. These cells migrate to the granule cell layer through the white matter [40]. Like granule cells, UBC are generated through a prolonged period of time, up to early postnatal stages [50].

Ventricular zone originating GABAergic interneurons, such as Basket cells, Golgi cells, and stellate cells arise from a pool of precursor cells and are produced from E13.5 till postnatal development [51–54].

Additionally to the neurons, the cerebellum contains various glial cell types, such as microglia, oligodendrocytes, Bergmann glia and parenchymal astrocytes.

2.3 Evolution

The cerebellum is found in all jawed vertebrates and its basic connectivity is thought to be conserved. Nevertheless, major differences in the development or organisation of the cerebellum can be found in various lineages. One of the most striking differences within the jawed vertebrate lineage is the presence or absence of a proliferative external granule cell layer. In birds and mammals, a transient structure emerges, surrounding the developing cerebellum, with proliferative granule cell progenitors. The presence of this group of cells is tighly regulated by SHH which is secreted by the underlying Purkinje cell layer [55]. In other lineages the presence of a proliferative EGL is not proven or debated [2, 3]. Whether the ancestral EGL had proliferative potential is under investigation [2]. Amphibians develop a non-proliferative external granule cell layer during metamorphosis which is suspected to help in distributing granule cells evenly throughout the cerebellum. This structure is also only transient in existence and its disappearance coincides with completed metamorphosis [2].

Zooming out from cellular development and complexity, the size of cerebella correlates directly with the size of the neocortex in mammals [56, 57]. A constant ratio of four cerebellar neurons to one cerebral cortex neuron was observed [56, 57]. This is an interesting observation, given that most evolutionary studies of mammalian brain development focused on the expansion of the neocortex. This cerebral structure is specific to mammals and exhibits the greatest increase in size in primates [58, 59]. Astonishingly, the ratio of cerebellar to cerebral volume seems to be elevated in great apes, assuming the previously established linear relationship [60, 61]. Most probably the increase in cerebellar size, co-occurring with the expansion of the cerebral cortex has its origin in the emergence of additional secondary progenitor zones, as recently shown for subventricular zone and rhombic lip progenitors in human [5]. Since the study by Haldipur *et al.* was mainly based on histological data, the molecular basis of the observed expansion is yet to be uncovered.

Another example of cerebellar specialization is the presence of the vermis, which is a region distinguished in mammals only. As mentioned initially, the emergence of the cerebellar anlage

is *FGF8* dependent [21]. The same gene regulates the appearance of the vermis in mammals, and additionally the isthmus. It is believed that the presence of *FGF8* is important to initiate the cerebellar primordia but the reduction of *FGF8* signalling is needed to allow cerebellar cell differentiation, though constant *FGF* expression is needed for vermis generation [2]. This illustrates the fine-tuned molecular mechanisms during cerebellar development, which can exhibit lineage-specific variations, and lead to different phenotypes in the adult.

The evolution of brain regions is thought to happen in a similar manner as the development of novel cell types and genes: using previously present predecessors and allowing them to change their function and behavior after duplication in another context. Recently, using cerebellar nuclei as a model, Kebschull *et al.* demonstrated that the multimerization of cerebellar deep nuclei from a single pair in cartilaginous fishes and amphibians, to two pairs in reptiles and birds, and three pairs in mammals followed the same principle [62]. This process involved the duplication of the full set of neuronal cell types which were present in the ancestor of the deep nuclei [62].

2.4 Studying evolution in developmental context

One of the fundamental questions biologists asked (and still ask) themselves is: how can the huge variety of phenotypes, even within similar or the same lineage be explained? After understanding the basics of genetics, this question grew to be an even greater mystery: many species share very similar genomes, yet the phenotype can establish itself as extremely different. One example, which goes back to the father of evolutionary theory, Charles Darwin, are the accordingly named Darwin finches. Even though genomic alterations are present between the species [63], the variety in beak phenotypes could not directly be understood. Studies showed that expression level differences of key genes can modulate the phenotype and therefore contribute to evolutionary processes [64].

Carroll wrote about the dilemma evolutionary biologists faced when studying genomic alterations in isolation: “The second major surprise was the similarity of proteins from species that looked and behaved as differently as, for example, chimps and humans. Mary-Claire King and Allan Wilson underscored the apparent paradox that presented and the challenge »to explain how species, which have such substantially similar genes can differ so substantially in anatomy. . . « (King and Wilson, 1975). They, like Zuckerkandl and others (Britten and Davidson, 1971), suggested that the evolution of anatomy occurred more by changing gene regulation than by changing protein sequences.” [8].

Nevertheless, it is not enough to compare gene expression profiles of adult tissues to understand the phenotypic evolution. The adult function, morphology and connectivity is established during development. Therefore, changes in developmental processes during evolution can have a major impact on the phenotypes in the adult [8]. One famous example of differential expression of key genes during development is the expression of *Shh* during snake development [65]. *Shh* expression in limb bud of snakes is not initiated due to the loss of an important enhancer region [65]. Without

studying Shh gene expression during snake and vertebrate development (figure 2 in [65]), this mechanism underlying the loss of limbs in the snake lineage, would probably have been missed.

This realization laid the groundwork for modern evolutionary and developmental (evo-devo) studies. [7, 66, 67]. Using bulk RNA-sequencing (RNA-seq) evo-devo data, gene expression in multiple tissues was studied by Cardoso-Moreira *et al.* [7]. The results of these studies confirmed multiple expectations evo-devo biologists had previously and added multiple additional questions which need to be answered by following studies. For example, the authors distinguished genes that are utilized during development in multiple organs (i.e. pleiotropic genes), and demonstrated that these genes are under higher evolutionary constraints than tissue-specific genes. This fits with the principles of evo-devo theory, as summarised in [8]. The bulk RNA-sequencing studies also identified genes that have evolved new expression trajectories during development in different lineages, but the molecular mechanisms underlying these changes remain to be elucidated.

2.5 Motivation and aims

Among the organs studied using bulk RNA-sequencing data [7], the cerebellum emerged as the one with the highest numbers of genes that show species-specific gene expression trajectory changes during development. Whether these changes occur due to gene regulatory alterations, or cell type composition changes could not directly be answered using the bulk RNA-seq data. This observation, the correlation between the numbers of neurons in the cerebellum and cerebral cortex during mammalian evolution, and the importance of the cerebellum in childhood carcinogenesis motivated the project in focus of this thesis. Thus, I aimed to characterise the development and evolution of the mammalian cerebellum, using cutting edge single-cell RNA-seq technologies that allow the dissection of gene expression at single-cell resolution.

3 Results

I investigated the development and evolution of the cerebellum in three mammalian species: human, mouse (*Mus musculus*) and opossum (*Monodelphis domestica*). Eutherian mammals (human and mouse) split from the marsupial lineage (opossum) approximately 160 million years ago, and the rodents from the human lineage 90 million years ago (figure 1). Single-nucleus RNA-sequencing (snRNA-seq) data was produced for a wide range of pre- and postnatal cerebellar samples. The tissue samples were prepared by Dr. Mari Sepp and if not otherwise mentioned, the whole developing cerebellum or its representative parts were used. This approach generated a detailed picture of the transcriptomic landscape of the developing and adult cerebellum for the studied species.

3.1 Dataset overview

This project was designed that it covers the development of the cerebellum in all three species with high resolution. Dr. Mari Sepp, Prof. Henrik Kaessmann and I decided to produce the data using the 10x Chromium 3' method. This approach is a droplet-based system which needs about 15,000 nuclei as input material and can generate about 5,000 nuclei per run. The nuclei isolation protocol was developed by Dr. Mari Sepp based on previously published methods [68].

Due to the strong differences in developmental pacing, especially between human and mouse, *a priori* estimate of agreement between the three timelines was needed. The work of Margarida Cardoso-Moreira *et al.* [7] guided the initial stage selection with the addition of earlier stages (down to embryonic day 10.5 (E10.5) in mouse). The project was designed to capture major landmarks of cerebellar development, including the birth and differentiation of the main cerebellar neuron types. Selected stages were: (I) E10.5, E11.5, E12.5, E13.5, E14.5, E15.5, E17.5, P0, P4, P7, P14 and adult (P63) in mouse; (II) 7 weeks post conception (wpc), 8 wpc, 9 wpc, 11 wpc, 17 wpc, 20 wpc, newborn, infant, toddler and adult in human; (III) E14.5, P1, P4, P5, P14, P21, P42, P60 and adult in opossum. For each stage, at least two biological replicates were included and for some individuals data from multiple runs (batches) were produced to increase the number of sequenced nuclei (summarized in figure 1). The raw data of each batch underwent quality control and barcode selection as described in section 5. In general, approximately 10,000 nuclei passed the filtering steps per stage per species (figure 1.A). Due to the known limitations in transcriptome coverage in single nuclei experiments, the distribution of UMIs (unique molecular identifiers) per cell was evaluated (figure 1.B). During data generation, 10x Genomics changed the available Chromium 3' kits from version 2 to version 3. The improved transcript capturing rate is clearly visible in the distributions of unique molecular identifiers (UMI) per cell: version 2 kits generated about 1,000, version 3 about 3,000 UMIs per nucleus. This difference in UMI counts is also reflected in the number of detected genes (figure 1). In total 395,736 nuclei from 87 libraries were sequenced and passed all quality

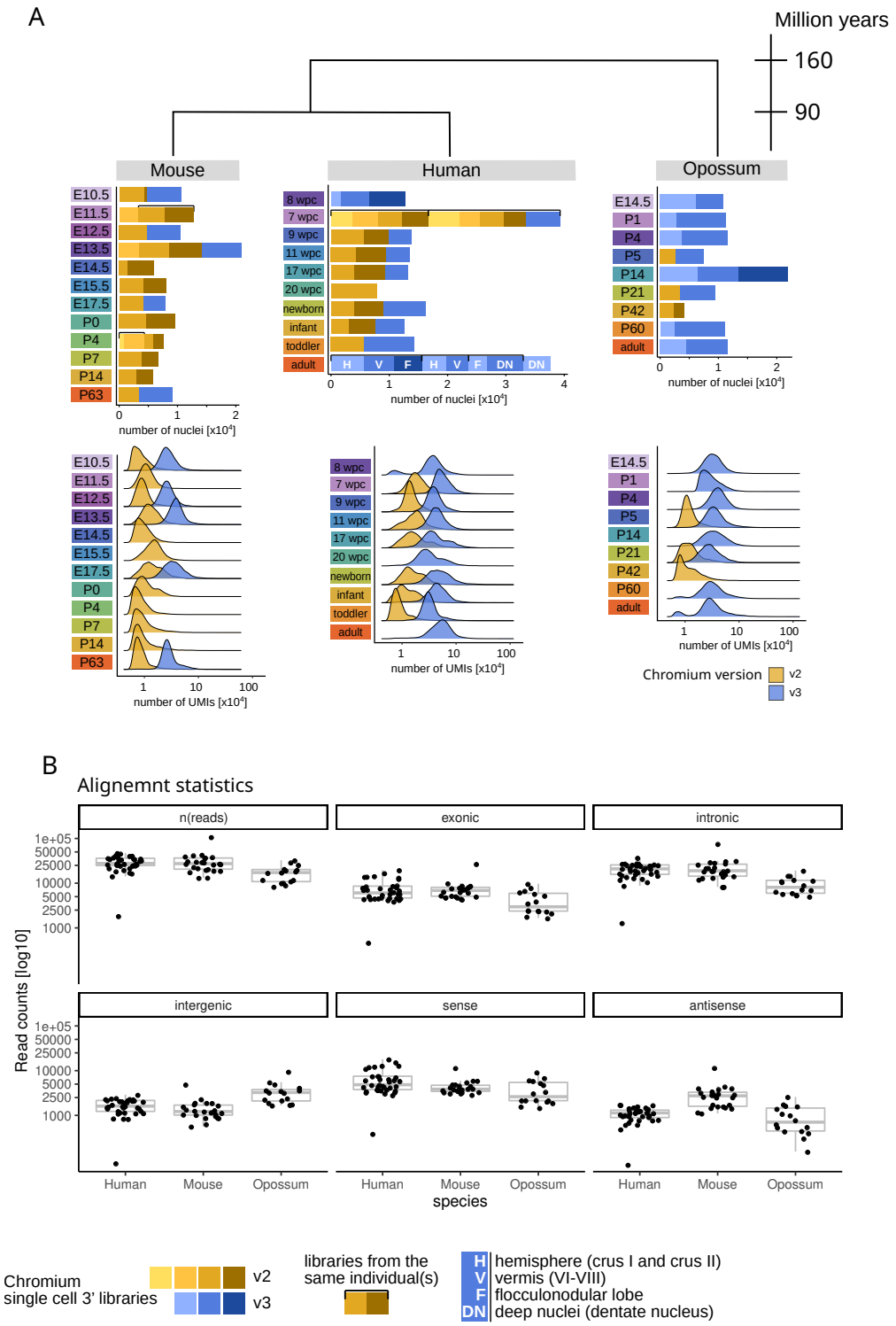


Figure 1: Dataset overview A: Phylogenetic tree (not to scale) of the three studied species. Upper barplots depict the total number of nuclei passing all quality control filters. Colors indicate 10x Chromium version (yellow shades: v2, blue shaded v3). Shading separates prepared libraries replicates. Black brackets enclose libraries prepared of the same individuals. For adult human the cerebellar regions of sample origin are indicated. The lower row of plots shows the distribution of unique molecular identifiers (UMI), grouped by the used 10x Chromium version, with log-scaled x-axis. **B**: Overall alignment statistics per library of the dataset. n[reads] = Number of reads sequenced; exonic = Number of reads aligned using exonic counting mode; intronic = Number of reads aligned using intronic counting mode; intergenic = Number of reads aligning in intergenic regions, exonic counting; sense = Number of reads aligning sense to the feature, exonic couting; antisense = Number of reads aligning antisense to the feature, exonic couting.

control filters with a median of 2,354 UMIs per cell across the studied species. For human 180,956, for mouse 115,282 and for opossum 99,498 nuclei were considered in the following analyses.

To assess the reproducibility of the generated data, I summed reads from each library in pseudobulks. Spearman correlations between pseudobulk profiles were calculated within each species using all genes that were expressed in at least 10% of all cells in any given sample (figure 2). The calculated species-internal correlation coeffients range from Spearmans -0.5 to 0.9 (figure 2). Clear aggregation of high agreement between samples from the same stage is visible (figure 2). Nevertheless, substantial drops in correlation within a stage are detectable, driven by the difference between Chromium 3' version 2 and version 3. The correlation coeffients are the highest between libraries generated from the same individual / pool of individuals. Taken together, these results demonstrate high quality of the datasets.

3.2 Initial data processing approach

During the phase of data generation, I explored various approaches to process and analyze the data. The main goal during initial data processing was the correct identification of informative genes and cell clusters in individual batch datasets. Prior steps of data processing and QC are are described in detail in the material and methods section (section 5). Compared to classical bulk RNA-seq experiments, the complexity of a single snRNA-seq library is already very high: the data includes approximately 5,000 nuclei and UMIs for about 25,000 genes. Due to technical drop-outs or biological variation (where the latter is probably the main source, see [69]), the majority of genes in single cell experiments have zero counts. To reduce the dimensionality of the data and therefore increase the signal to noise ratio, first the data was normalized by sampling depth (sum of UMIs), and then the highly variable genes (HVG) were called. For the identification of informative genes (i.e. HVG) I used an approach that was originally developed by Prof. Simon Anders. The approach is based on the hypothesis that uninformative genes follow a Poisson distribution: a gene has the same probability of detection (generating a UMI) in all sampled cells. This means that the mean UMI count, μ, is equal to the variance of the given gene, σ^2. If, however, a gene is specifically detectable in a subset of cells, for example being cell type specific, or at least higher expressed in a group of cells, the gene will exhibit a Poisson mixed model, which means that $\sigma^2 > \mu$. Furthermore, due to differential sequencing depth between cells, normalization is needed prior to further analyses and modeling. Per gene, its mean and variance relationship (VMR) is calculated and compared to the Poisson expectation ($\Xi = \frac{1}{N}\sum_j \frac{1}{s_j}$). If the VMR exceeds Ξ, a mixed Poisson distribution is assumed and the gene is regarded as informative.

I chose this approach, even though other, similar methods exist, for example Seurat [70]. The reasons for my decision were the easily understandable statistical basis and the possibility to determine the number of HVGs by selecting a factor for Ξ, thus allowing the number of HVG to be

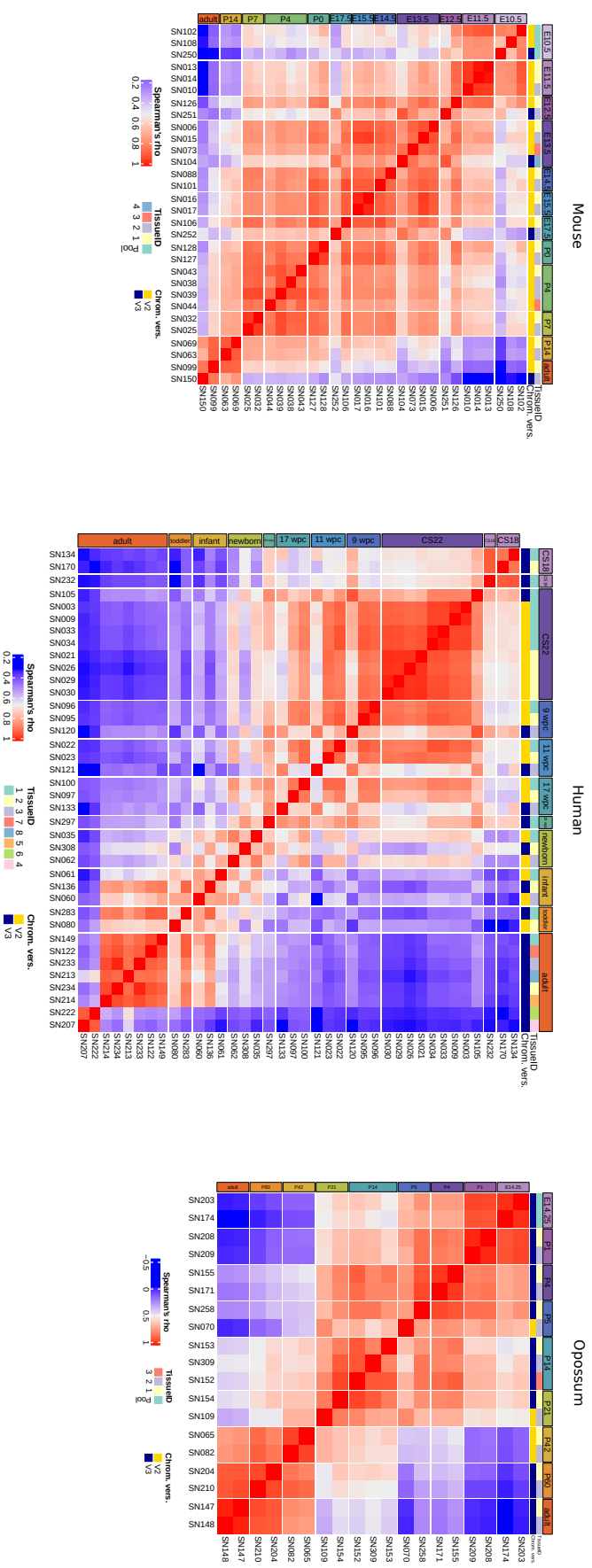


Figure 2: Correlations across libraries in mouse, human and opossum datasets Heatmaps of pairwise Spearman correlation coefficients per species and pseudobulk. Pseudobulks were generated by summing up all UMI per library. Biological replicates and 10x Chromium versions are indicated using colors above the columns.

informed by the data itself rather than assuming a specific number. For comparison, in Seurat it is required to set a fixed number of HVGs.

Following normalization and selection of HVGs, further dimensional reduction was accomplished by applying principal component analysis (PCA) to the normalized and scaled HVG by cell matrix. An elbow plot, visualizing the variance explained by each principal component (PC) could be used to determine the number of PCs to keep, but anecdotally, I observed that keeping more components less strongly influences the final uniform manifold approximation and projection (UMAP) and clustering than removing too many components. Henceforth, I decided to keep at least 25 components per sample.

For accessible data visualization and summarization, the UMAP [71] algorithm was applied to the principal component embedding. UMAP is the de facto standard for single cell data visualization and the R implementation (uwot [72]) provides easy access to the approximated neighborhood graph, which is used as an imput for some of the downstream analyses.

After these initial preprocessing steps, cells need to be grouped into meaningful clusters due to the low sequencing depth of individual cells. Whether the called clusters reflect individual cell types need to be explored using literature and other available data resources. We opted for the Louvain clustering [73] as implemented in the Python scanpy package [74]. As with UMAP, the Louvain algorithm is the default in many single cell analyses. Since the clustering should only give an initial approximation at the possible groupings, and variation of the dataset, annotation based on previous literature on cerebellar cell types will allow post-hoc merging of similar clusters.

All of the aforementioned steps, done for individual samples, allowed evaluation and establishment of data processing methods. Given that the presented study includes 87 snRNA-seq libraries from 78 independent samples, batch correction and data integration were needed for further analyses.

3.3 Batch correction and data integration

To take batch effects into account is needed in most single cell studies which include more than one biological replicate, due to the delicate technology and strong reduction of dimensionality. Many methods are leveraging a k-nearest neighborhood to model the high dimensional manifold, such as UMAP and Louvain clusters. If individual biological or technical replicates show differential gene expression, these neighborhoods are called within each batch and, depending on the size of the dataset, the k-nearest neighborhood graph is partitioned by batch due to the limited number of neighbors (i.e. number of edges in the graph) considered.

The easiest methods to remove these influences include linear regression to model the confounding factor, others use more complex statistical approaches to model technical artifacts [70, 75–77]. If the correction algorithm is too greedy in pushing the datasets on top of each other, true biolog-

ical signals can be masked or severely distorted. The datasets presented here span a whole organs development and therefore, differences between stages are expected and are biological meaningful. On the other hand, the same cell type which is present in multiple stages, should exhibit a distinct signal which should be recoverable to discriminate one cell type from the others. Finding the balance between integration across stages and not losing stage-specific signals was the main goal. I tried various methods, but the results of these tests shall not be part of this thesis. The conclusion was that LIGER [77] performed best in merging without over-integration of the datasets, judged by naive UMAP inspection and gene enrichment analyses. It was not only able to integrate data from batches of single stages (figures S1,S2,S3), but also to construct a continuous integration across stages (figure 3). Nonetheless, especially in the human dataset, strong deviations of the embedding, correlating with the batch (or biological replicate), are detectable (see figure S2 facet: newborn).

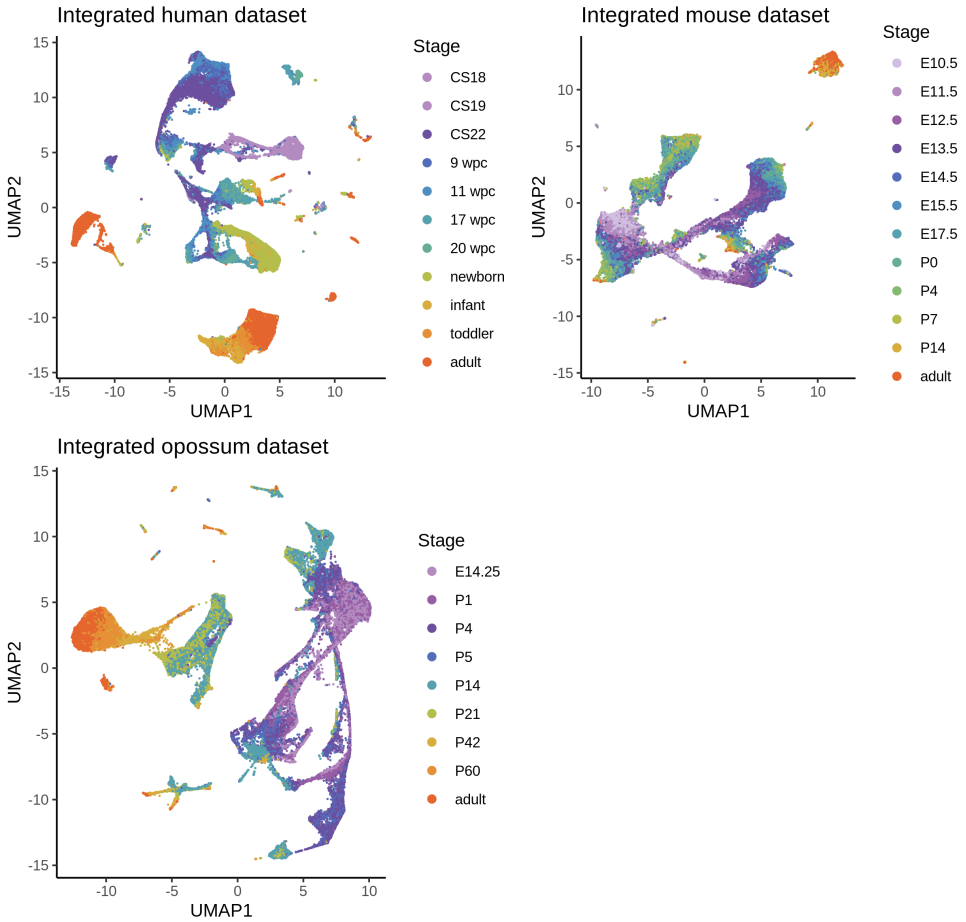


Figure 3: UMAP embeddings of the human, mouse and opossum datasets integrated by LIGER Nuclei are colored according to the stage they were sampled from. Colors reflect aligned stages (according to the results in the following chapter).

Even though the different stages overlap substantially after the LIGER integration, individual areas of the UMAP enrich for specific stages. This is very pronounced for the postnatal and adult stages, which are present in very specific clusters without much pre-natal cell contribution. The occurrence of individual stages within the UMAP are often in chronological order, meaning,

for example, cells of stage E13.5 in mouse are flanked by cells of the adjacent E12.5 and E14.5, indicating a preservation of a developmental signal, despite the batch correction. Additionally, this preservation of the developmental signal in the datasets of all three species results in a trajectory, like appearance of multiple broad cell type lineages in the integrated UMAPs .

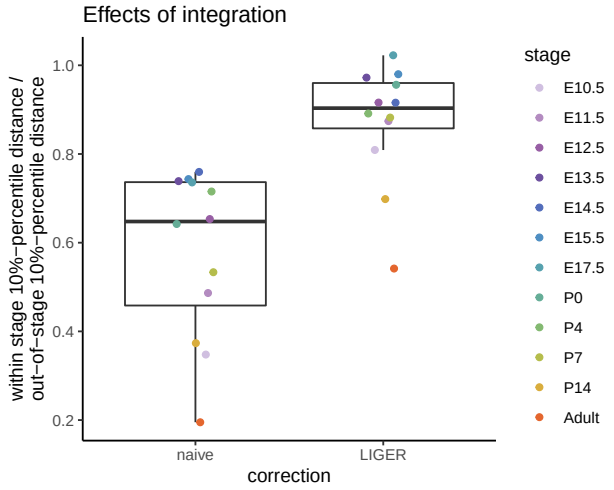


Figure 4: Effect of integration on distance values To illustrate the effect of batch correction, using LIGER correction as example, I calculated the 10%-percentile within stage distance and the same percentile interstage distance, using the mouse dataset. The quotient of both values is calculated per stage. The naive calculation was based on a PCA embedding, based on normalized expression matrices and the LIGER values were calculated based on a 100-dimensional LIGER non-negative matrix factorization result.

To investigate the effect of integration on the interpretability of the low dimensional embeddings of the data, I quantified the within-stage and across-stage Euclidean distance relationships: I first naively processed all cells without any batch correction by HVG selection, normalization and PCA. After that, I used the pan-stage LIGER embedding to retrieve the same values. I extracted for each stage the 10%-percentile distance within stage and to all other stages. This was done for the PCA and for the LIGER factors. Finally, I calculated the ratio of within-stage Euclidean distance to out-of-stage distance (figure 4). For the naive PCA approach, the median ratio was 0.65 and for the LIGER integration approximately 0.9. If the PCA dimensional reduction can be assumed to reflect the transcriptomic landscape *ab initio*, the difference of the within- to out-of-stage 10%-percentile distance captures the shift in transcriptomic space during development. Once LIGER was applied, the difference shrinks substantially, facilitating cell type clustering (see below) but underestimating stage-specific expression differences. This distortion is captured in the UMAP visualizations as well, and therefore needs to be kept in mind when interpreting the low dimensional embeddings.

3.4 Mouse data annotation

When Dr. Mari Sepp and I started to annotate the cell types in our snRNA-seq datasets, we knew that the majority of literature sources are based on mouse development. Thus, we decided to annotate the mouse dataset first and use the gained knowledge and cell type associations for transfer

to the other two species. Transferred annotations were then set to be validated with external sources and checked for biological relevance to circumvent spurious matches that result in wrong conclusions.

As mentioned before, single cell RNA-seq data is very noisy and individual cell expression profiles lack the stringency to call presence or absence of a particular transcript or the expression levels. Hence, I clustered the datasets using the aforementioned Louvain algorithm with high resolution into 61 (mouse), 68 (human) and 67 (opossum) clusters. Since these clusters were called on the merged datasets, I assumed that even this high number of clusters does not fully capture the complexity of the datasets. To improve the resolution, I integrated the data separately for each cluster using LIGER, and called subclusters. This iterative clustering approach resulted in approximately 590 clusters per species. It is important to note here, that I did not assume that there are 590 cell types or states present in the dataset, but this very granular clustering allowed me to characterize each group of cells and merging of (sub)clusters, wherever needed.

Using a term-frequency inversed document frequency (TF-IDF) transformation for gene scoring and a hypergeometric test for p-value estimation, for each cluster significantly enriched ($p < 0.01$) genes were called. The called genes were used by Dr. Mari Sepp and me to annotate the mouse clusters using literature [4, 78] and freely available data repositories, such as the Allen mouse developing brain atlas [79, 80] and GenePaint [81].

The high complexity of the dataset made it necessary to apply a hierarchical annotation strategy (figure 5):

(I) Based on developmental origin, we grouped cells into broad cell type lineages. The VZ lineage includes GABAergic neurons born at the cerebellar ventricular zone; RL/NTZ comprises early-born rhombic lip-derived glutamatergic neurons that assemble at the nuclear transitory zone; RL/EGL includes neurons originating from the late rhombic lip that is associated with a secondary germinal zone in the external granule cell layer. The remaining groups are glia, cells of mesodermal origin, and neural cells from neighbouring brain regions (other).

(III) We defined cell states as groups of cells sharing the same cell type and the same level of differentiation, no matter whether they originated from the same developmental stage (e.g., granule cell (GC) states are distributed over multiple stages). For instance, for the granule cells, we differentiated GC progenitors (GCP), differentiating GCs (GC_{diff1} and GC_{diff2}), and defined GCs. Similarly, the astroglia cell type includes cell states of neural progenitors, glioblasts and astrocytes.

(IV) We further divided cell types at some cell states into subtypes. This depended on the remaining variability and the number of cells collected. Examples of subtype categories include "$progenitor_{RL}$", "$progenitor_{VZ}$", and "$progenitor_{gliogenic}$".

All labels used at different levels of annotation are summarised in table 10. Altgothere, we specified 4 broad lineages, 26 cell types, 44 cell states and 48 subtypes. This approach allowed analyses at different states of cell type differentiation and, down the line, a more detailed comparison

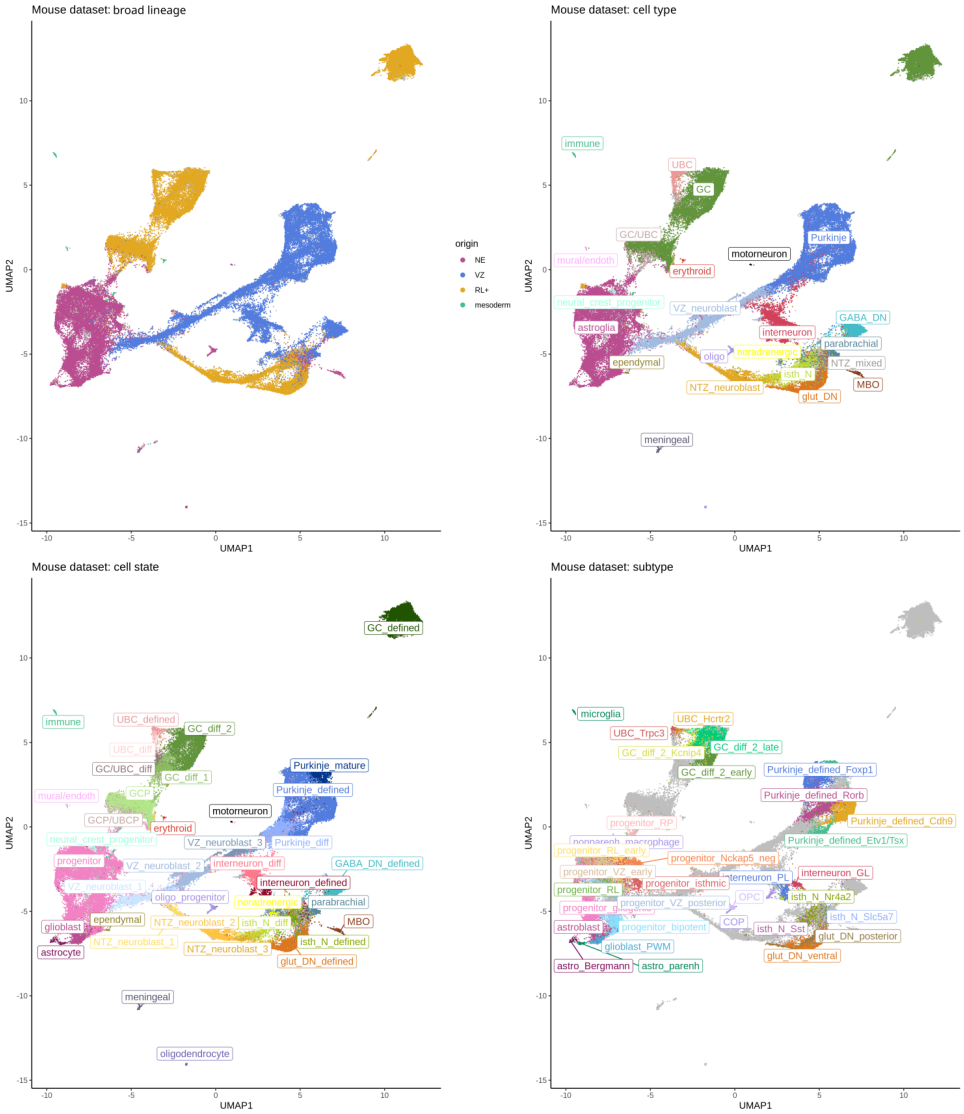


Figure 5: Mouse dataset annotation To compile a detailed annotation of the dataset, nuclei were annotated at different levels: upper left - Broad lineage, indicates the major developmental lineages. Upper right - Cell type, broad cell type which could be associated to each cell. Lower left - Cell types were split into cell states. This annotation level describes transcriptomic states of differentiation. Lower right - If residual variance was detectable, cells were labelled according the identified subtypes. Often these subtypes are named according to one major marker gene. Cells lacking a subtype label are depicted in grey.

of the human, mouse and opossum datasets.

3.5 Cross-species integration and annotation transfer

Once the mouse dataset was annotated, I integrated the human and opossum dataset separately with the mouse dataset. Since at the time of integration, most popular integration methods did not work with datasets the size and complexity of the one here presented, I needed to develop a custom approach to transfer the highly detailed mouse annotation to the human and opossum datasets. I settled on a two-step approach, which was then used to transfer the mouse annotation to the other species. First, I ran LIGER to project two species in the same 100 dimensional embedding . I only provided batch information and not the species information to the algorithm, using shared (between all batches of both species) detectable one-to-one orthologs in pre-mRNA counting mode (mouse with human = 6101 orthologs, mouse with opossum = 5019 orthologs). Previous tests informed me (data not shown), that the species label increased the runtime and did not improve the integration overall. In this initial projection, stage effects were reduced but the two species were still separate from each other. To further correct the species signal, I then applied MNN-correct [75] to the 100 dimensional embedding, only providing the species label. I then generated a UMAP, based on the MNN-corrected LIGER embedding. In this UMAP the two species shared a common structure and appeared overall merged (figure S4). I concluded that the correction was successful.

Now, that I had generated pairwise the corrected 100 dimensional embeddings for human and mouse as well as for opossum and mouse, I grouped the cells based on their Louvain subclusters, defined in the single species embeddings. For each subcluster I calculated the centroid in the corrected 100-dimensional pairwise embedding. I used the subcluster to component centroid matrices of both species and calculated a human to mouse and opossum to mouse correlation matrices (Pearson correlation). This matrix was then used as a guide to assign each human and opossum subcluster to the highest correlating mouse subcluster. As the mouse subclusters were linked with different level annotation labels (as described above), each human and opossum subcluster was associated with the corresponding annotation from the mouse and correlation coefficients were used as a confidence measure. To account for potential differences in sampling, all transferred labels were then scrutinized and if needed subclusters were re-annotated. This allowed the identification of cell types and states that were not captured in the mouse dataset, helping to get an unbiased view on cerebellar development and cell type composition in human, mouse and opossum (figure 8.A).

During the annotation transfer process, it became apparent, that one of the human samples (SN296) contained many cells that showed *HOX* gene expression, which is not expected to be expressed in the cerebellum, but in adjacent hindbrain regions. We concluded that the sample had not been correctly dissected and henceforward removed this sample from further analyses.

Additionally, I integrated all three species in one common embedding using the same approach

as described above. This full embedding was used for an analysis, inspired by La Manno et al. [82], which aims to create a continuous stage vector out of the discrete stage assignments via neighborhood grouping, called pseudoage. I modified the pseudoage calling by letting each human and opossum cell call its k-nearest neighbor in the mouse dataset within the common embedding. Using this information the human and opossum cells were assigned to a mouse corresponding pseudoage.

3.6 Stage correspondence calling

To compare human, mouse, and opossum over the vastly different developmental time-periods, the sampled discrete stages need to be paired. Human prenatal development lasts for approximately nine months, mouse prenatal development spans roughly 20 days, and opossum embryos are born at a very immature stage that is comparable to human six week embryos [83]. In all three species, neurogenesis in the cerebellum is ongoing for several weeks (mouse and opossum) or years (human) after birth (see introduction, section 1). I aimed to robustly match the sampled developmental stages across the three species based on three measures: (I) transcriptomic correlation, (II) pseudoage agreement and (III) cell state composition (figure 6). Dr. Mari Sepp and Ioannis Sarropoulos gave major input to this analysis.

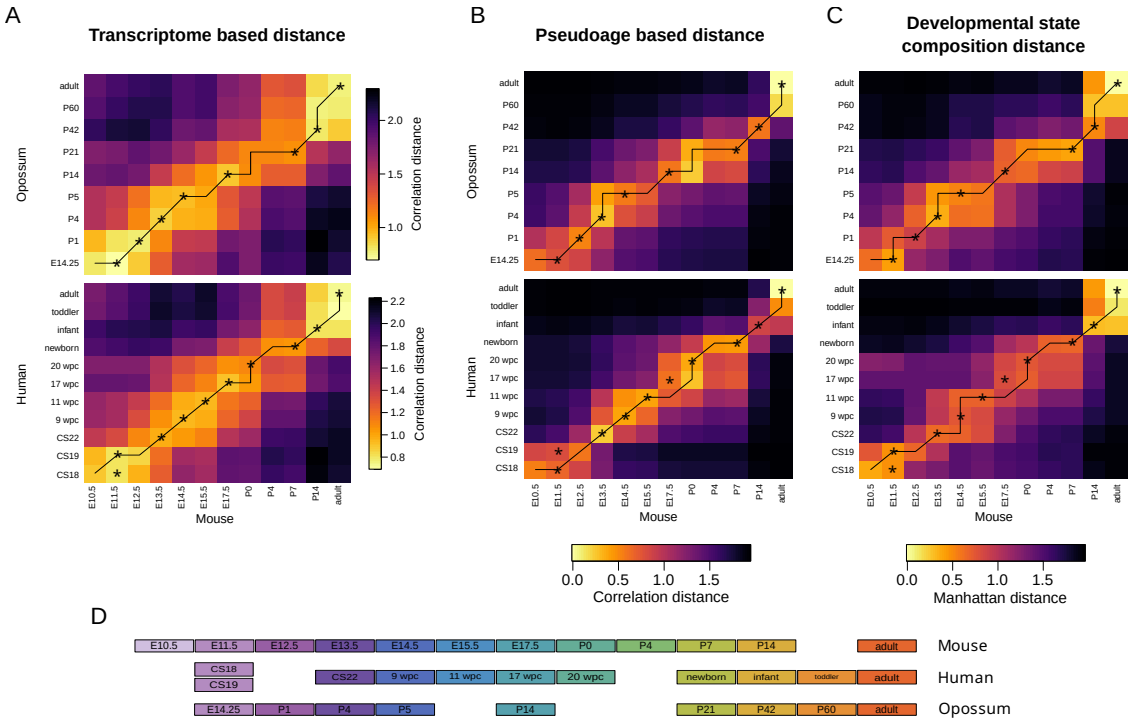


Figure 6: Stage correspondences A: Spearman correlation distances between developmental stages, calculated on normalized expression matrices, based on all detectable one-to-one orthologous genes. **B**: Manhattan distances between developmental stages, calculated using proportions of pseudoage assignments. **C**: Manhattan distances between developmental stages, calculated using proportions of developmental state (cell state) assignments. **D**: Aligned stage assignments. Colors reflect the alignment vector between all three species. Gaps show non-matched stages. The lines within each heatmap are the result of dynamic timewarp to determine the shortest path through the correlation matrices. Tiles with stars indicate the final decisions of stage correspondences.

I called transcriptomic correlations by aggregating all cells from the same developmental stage into a pseudobulk. To estimate pairwise Spearman correlation distances between the stage-specific pseudobulks, I considered all pairwise one-to-one orthologous genes that were detected in both compared species and called as highly variable in at least one of the species (figure 6.A). To find the pseudoage distances, I used the pseudoage assignments of the cells (see above), calculated the proportional pseudoage abundances for each developmental stage, and estimated the pairwise Manhattan distances between the pseudoage abundances (figure 6.B). Finally, to compare cell sate compositions, I determined Manhattan distances between the relative cell state abundances of the developmental stages (figure 6.C). For all three measures, dynamic timewarping was applied to identify the best alignment sequence between human and mouse or opossum and mouse (figure 6.D). Overall, the three measures agreed. In a few cases, where slight shifts in the assignments were detected (e.g., human 17 wpc assigned to E17.5 or P0 by the different measures), the consensus stage assignment was made based on the smallest transcriptome distance. Based on this analysis, we assigned, for example, the following correspondences: mouse E11.5 corresponds to human CS18 and CS19 and opossum E14.5, Mouse E12.5 is unmatched in the human data and matches with opossum P1. Mouse E15.5 matches with human 11 wpc but does not have a correspondence in the opossum dataset. Mouse E17.5 matches with human 17 wpc and opossum P14. The first mouse postnatal sample, P0, matches to the last human embryonic sample at 20 wpc and is not matching to any opossum sample. Mouse P7 matches to human newborn and opossum P21. Mouse P14 matches to human infant and opossum P42. Toddler in human and P60 in opossum which was not matched to mouse was called “intermediate”.

Throughout this thesis, when I mention matched stages or aligned staging, I am referring to the correspondences assigned in this analysis.

Overall the assigned stage correspondences, which I identified in this study, match to our assumptions, derived form Cardoso-Moreira et al. [7] (figure 7). As the stage matching in Cardoso-Moreira et al. is based on all somatic tissues and these assignements are in agreement with the assignments based on cerebellar snRNA-seq data of this project, I conclude that there is no detectable heterochrony in cerebellar development between the studied mammalian species.

3.7 Atlas of cell types and states

3.7.1 Overview of identified cell types

All previous analyses and approaches were done with the aim to compare the cerebellar development in the three species in as detailed manner as possible. In this chapter, I describe ... Gene names mentioned refer to the human ortholog, as long as the homology is unambiguous to the other species. Enriched genes listed below were called using a combination of TF-IDF transformation and

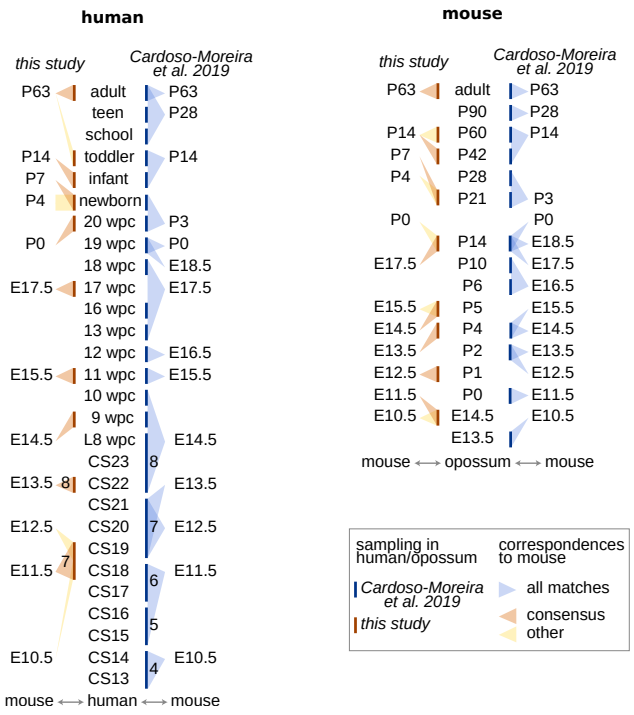


Figure 7: Stage matching comparison to bulk RNA-seq of somatic tissues [7] The illustration was prepared by Dr. Mari Sepp.

hypergeometric test for p-value determination.

Assuming that the integrated UMAPs of each of the three species reflect the overall cell type relationships correctly (which is rarely the case without major limitations [84]), mammalian cerebellar development is characterized by a star-like shape in transcriptomic space. Proliferative progenitor cells, as judged by cell cycle scoring (figure 8.B), are located centrally with differentiating cells of different lineages spreading out. At the cell type level, the progenitor cells together with the glioblasts and astrocytes (*SLC1A3* and *AQP4*) form the astroglia lineage. A clear separation between cycling ventricular zone progenitors (*KIRREL2*) and rhombic lip progenitors (*SLIT2*, *LMX1A*) is not detectable in any species, instead the progenitor cells are arranged along a continuum that reflects their position in space and time. The three major arms, originating in progenitor cells, represent the broad lineages of cerebellar neurons. RL/NTZ lineage comprises cells that differentiate into glutamatergic deep nuclei neurons and extra-cerebellar isthmic nuclei neurons. RL/EGL lineage gives rise to granule cells (*PAX6*, *GABRA6*) and unipolar brush cells (*LMX1A*). VZ lineage gives rise to, depending on the stage, parabrachial neurons (*LMX1A* and *LMX1B*), noradrenergic neurons (*LMX1B* and *PHOX2B*), GABAergic deep nuclei neurons(*SOX14*), Purkinje cells (*SKOR2*), and GABAergic interneurons (*PAX2*). Parabrachial and noradrenergic neurons will migrate out of the cerebellum to the brainstem, later in development [85].

Oligodendrocyte differentiation could also be observed in all three datasets. We detected oligodendrocyte progenitors (OPC, *PDGFRA*), committed oligodendrocyte precursors (*TNR*) as well as postmitotic oligodendrocytes (*MAG*). In human and opossum, a group of cells that resembles

a transitory state between astroglial progenitors and OPCs could be singled out (*EGFR*), which likely represent the previously described preOPCs [82]. The abundance of preOPCs is high in the opossum dataset but very low in the human dataset Ependymal cells (*SPAG17*) were readily detectable in mouse and opossum, but were lacking in the human dataset (likely as a result of sampling differences). Some extracerebellar cell types were also detected: cells expressing *LEF1*, which reside in the adult cerebellum but are born in the midbrain, and cells that can be considered contaminations from adjacent brain regions, such as motor neurons, isthmic neuroblasts, midbrain neuroblasts and GABAergic midbrain cells. Finally, typical non-neuronal cells could also be found: mural and endothelial vascular cells, erythroid and immune cells (the majority of which were microglia) and meningeal cells. A general pattern of cell cycle score reduction during development could be observed in all cycling cell types (figure 8.C)

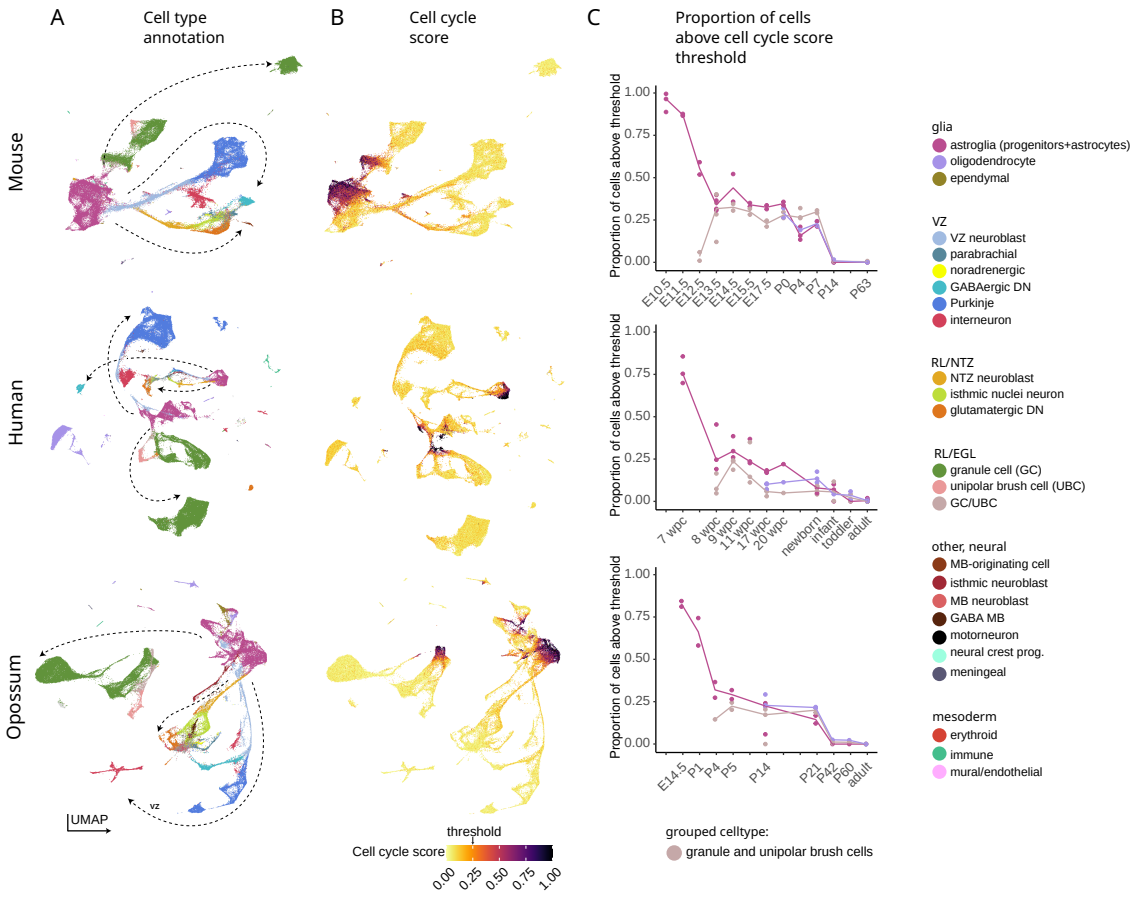


Figure 8: Cell type level annotations and cell cycle scores A: Cell type assignments for mouse, human and opossum datasets. Arrows indicate differentiation trajectories for easier visual comparison **B**: Cell cycle score per nucleus per species. Darker colors indicate higher cell cycle score. This score is an approximation of the cell cycle activity (higher = more cell cycle activity). **C**: Cell cycle scores were thresholded hat 0.25 and proportion of cycling cells (score > 0.25) per cell type was determined.

3.7.2 Cellular dynamics

Having a high resolution picture of cell types in the human, mouse and opossum cerebellum next to a refined stage matching scheme, allowed me to compare the cellular composition dynamics during cerebellar development across the mammalian species. I conducted this analysis using the "cell type" annotation level to have a robust foundation for comparisons. As expected, the mammalian cerebellar development is a highly dynamic process with strong shifts in relative cellular abundances. All three species share a common pattern (figure 9.A). Early in development, the majority of cells are cycling progenitors, making up 50 - 70 % of all cells. Next, there is a phase of GABAergic cell expansion (mainly VZ neuroblasts and Purkinje cells) and these cells dominate the cell type composition up to around mouse stage E15.5. The last GABAergic cells to appear are interneurons. After this wave, granule cells take over the cell type landscape: the external granule cell layer is established already at mouse embryonic stage E13.5, but the majority of granule cells are produced at later stages. The strong amplification of GCs in EGL results in granule cells making up more than 80% of the adult cerebellar cells. For this analysis it is important to note, that human samples up to 11 wpc should be representative of cell type proportions, but after that, only pieces of the human cerebellum could be sampled, possibly leading to biases in the estimation of relative cell type abundances.

Besides these similarities in overall cell type dynamics, differences could be recognized: during two consecutive matched timepoints (stages matched to E13.5 and E14.5 in mouse), Purkinje cells have twice as high relative abundances in human as compared to the other two species (60% vs 30%) (figure 9.B). To investigate whether this difference is supported by statistical methods, I created a Bayesian hierarchical model to capture the observed cell type proportion measures: I assumed that cell type proportions in each stage, batch, and species follows a binomial distribution, where the proportion is represented as the probability p_b, the total number of cells per cell type, batch, stage, and species as y_b, and the total number of cells as N_b (equation 1).

p_b is assumed to be sampled from a species and stage wide hidden normal distribution with standard deviation σ^2. The mean of this distribution, p_0, is sampled from a Students-T distribution. This was chosen to make the mean of the per species and stage normal distribution more resilient against outliers. The goal was the estimation of hyperparameter p_0 to compare the expected proportion of cells per cell type, stage and species to each other. As I modelled all stages and species at the same time, I sampled from the posterior of p_0 and estimated the posterior difference (figure 9.B).

$$
\begin{aligned}
y_b &= Binomial(N_b, p_b) \\
p_b &= Normal(p_0, \sigma^2) \\
\sigma^2 &= Exponential(1) \\
p_0 &= StudentsT(1, 1.5, 1)
\end{aligned}
\tag{1}
$$

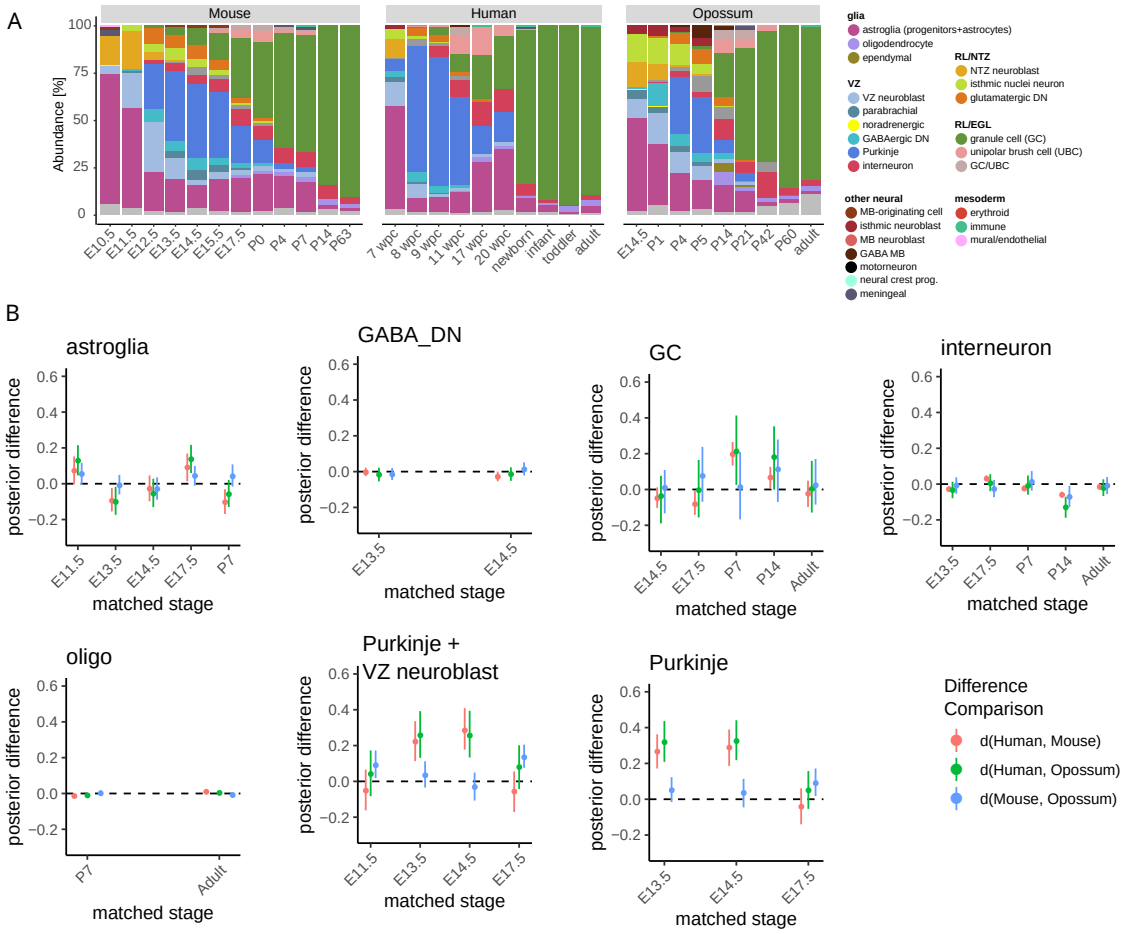


Figure 9: Bayes modelling of differences in cell type relative abundances A: Proportions per cell type per developmental stage per species. Cells not assigned to a cell type are in grey **B**: Posteriors of Bayesian model, modelling the pairwise difference of cell type abundances. The point represents the mean, the line range represents the 95%-highest density interval (HDI) of the posterior distribution. Zero difference is indicated with a dotted black line. This plot can be used for region of plausible equality (ROPE) estimation: when the HDI intersects the zero difference line, plausible equality is assumed. Pairwise comparisons between species are shown in different colors. The model was built for each matched stage (x-axis) and labelled according to the mouse stage.

The majority of cell types, which were modelled, showed no strong cell type overrepresentation at specific timepoints between the species. Here, it is important to note, again, that comparability between all three species is only given up to matched timepoint E15.5. Slight increases in the abundance of human astroglia cells could be detected at matched timepoints E11.5 and E17.5, which were approximately 10% above the abundances in mouse and opossum. The most striking difference was detected for Purkinje cells in human, as mentioned before. The highest density interval (HDI,

mass = 95%) indicates statistical evidence, that the observed difference can be supported by the chosen model. To circumvent possible variations in the blurry boundary between VZ neuroblasts and Purkinje cells, another model was fit to a group of cells that combines Purkinje cells with all VZ neuroblasts. Again, the HDI shows no intersection with zero difference, hence the model supports the presence of increased abundances of developing Purkinje cells in human.

3.7.3 Global gene expression patterns

Next, my goal was the characterization of the global gene expression patterns during cerebellar development.

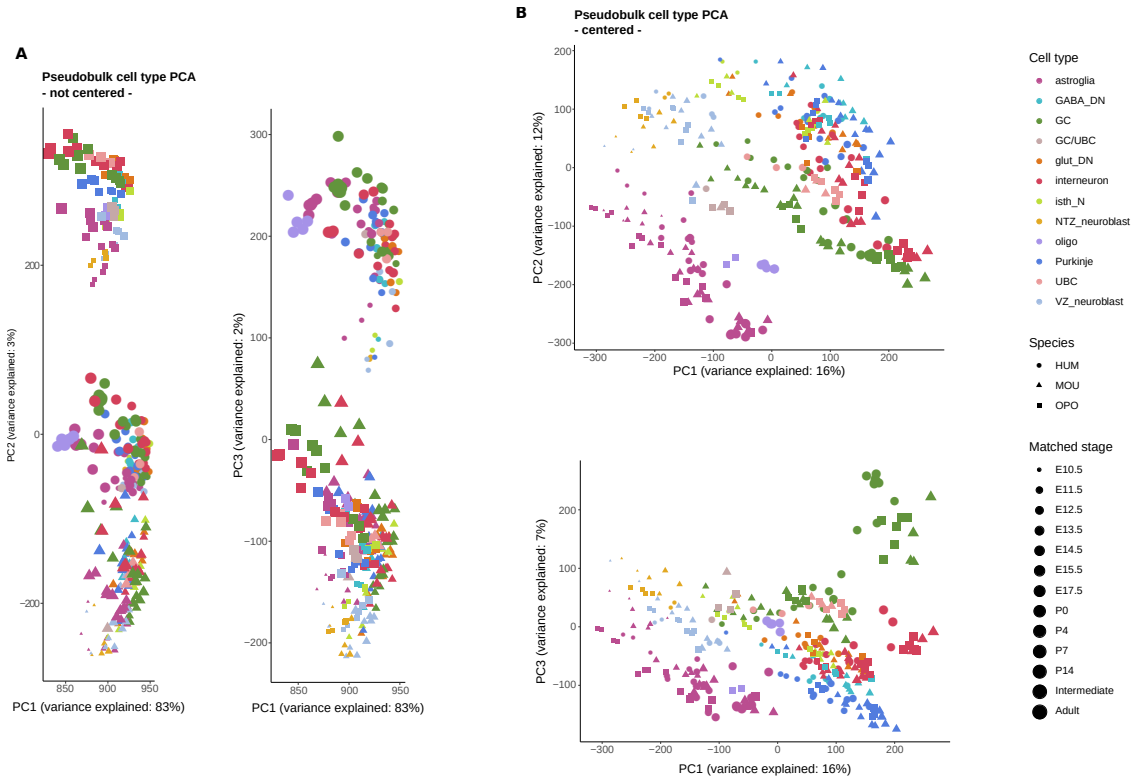


Figure 10: Principal component analysis (PCA) of cell type pseudobulk transcriptomes **A**: First three principal components (PC) of non-median centered expression matrix. Cell types are color-coded and species shown in different shapes. **B**: First three PCs of median centered expression matrix. Both PCA were generated using three-way one-to-one orthologs (n = 10,276) and exonic UMI.

I created cell type-specific pseudobulks for each biological replicate, mimicking bulk RNA-seq experiments. Next, I subsetted the genes for one-to-one orthologous genes between all three species (10,276 genes, exonic UMI counts). Inspired by an analysis by Cardoso-Moreira et al. [7], I ran a principal component analysis on the combined dataset to assess the broad transcriptional landscape and to investigate the sources of variation (figure 10). According to the variation explained by the first 3 principal components, the main sources of variation are independent of the species (PC1, explains 83% of variance), and less variation is associated with species-specific signals: PC2 explains 3% of variance and separates opossum from therian mammals, PC3 explains 2% of variance and

separates human from the other two species (figure 10.A). I decided to center the expression values per gene and species to given that centering enables the PCA to capture relative expression patterns in each species, and ignore absolute expression differences, which might arise due to technical artifacts (e.g., differences in sequencing depth, Chromium version). The first two components of the centered PCA show clear developmental and cell type specific signals, the separation of neuronal from glial cell types is especially evident (figure 10.B). The third component further separates the individual neuronal cell types (figure 10.B). Pseudobulks of early stages cluster closer together and clusters widen up when cell type differentiation progresses (PC1 and PC2). Overall the first three components of the centered PCA capture 35% of variance, compared to 88% in case of the not-centered input. General visual inspection of the PCA shows strong similarities to the single cell UMAPs (compare 10.A and figure 8.A).

Altgother, these analyses indicate that developmental and cell type signals explain the majority of gene expression variance in the developing cerebella from the three mammalian species.

3.7.4 Conserved and divergent cell state markers

The eutherian and marsupial species are separated by at least 160 million years of mammalian evolution. The above PCA indicated, that the core gene programs of relative gene expression are shared between the three studied species. Hence, Dr. Mari Sepp and me speculated that genes which have cell state-specific expression pattrens that are conserved in all three studied species, are likely of high importance in defining the cerebellar cell types during development. I tested whether a given gene is enriched in a given cell state, compared to all other cell states, combined. To ensure comparability between the species, only cell states and stages which were sampled in all species were considered in this analysis. TF-IDF transformation was done, followed by hypergeometric test. The data was reduced to only contain one-to-one orthologs. An illustration of species-specific and conserved marker gene scaled expression values is shown in figure 11.

In each species approximately 61% of all identified markers are specific to a single cell state, followed by genes which are enriched in two cell states (on average 26%) and genes which are enriched in three or more cell states (averaging to 13%). For overlapping marker genes (called as markers in multiple cell states), only a minority (approximately 10%) is called for cell states of the same cell type (data not shown).

Additionally, the majority of cell state markers, amongst the one-to-one orthologs were only called in a single species (figure 12). Various combinations of shared expression pattern were also observed, like enrichment in human and mouse and a lack thereof in opossum (figures 12, S7).

A general characterization of the conserved cell state marker genes (examples in figure 14) was done by enrichment analysis ($p < 0.01$, empirical enrichment > 2) of associated gene onthology (GO) terms. Progenitor conserved marker genes enrich in cell cycle associated GO-terms like “chromo-

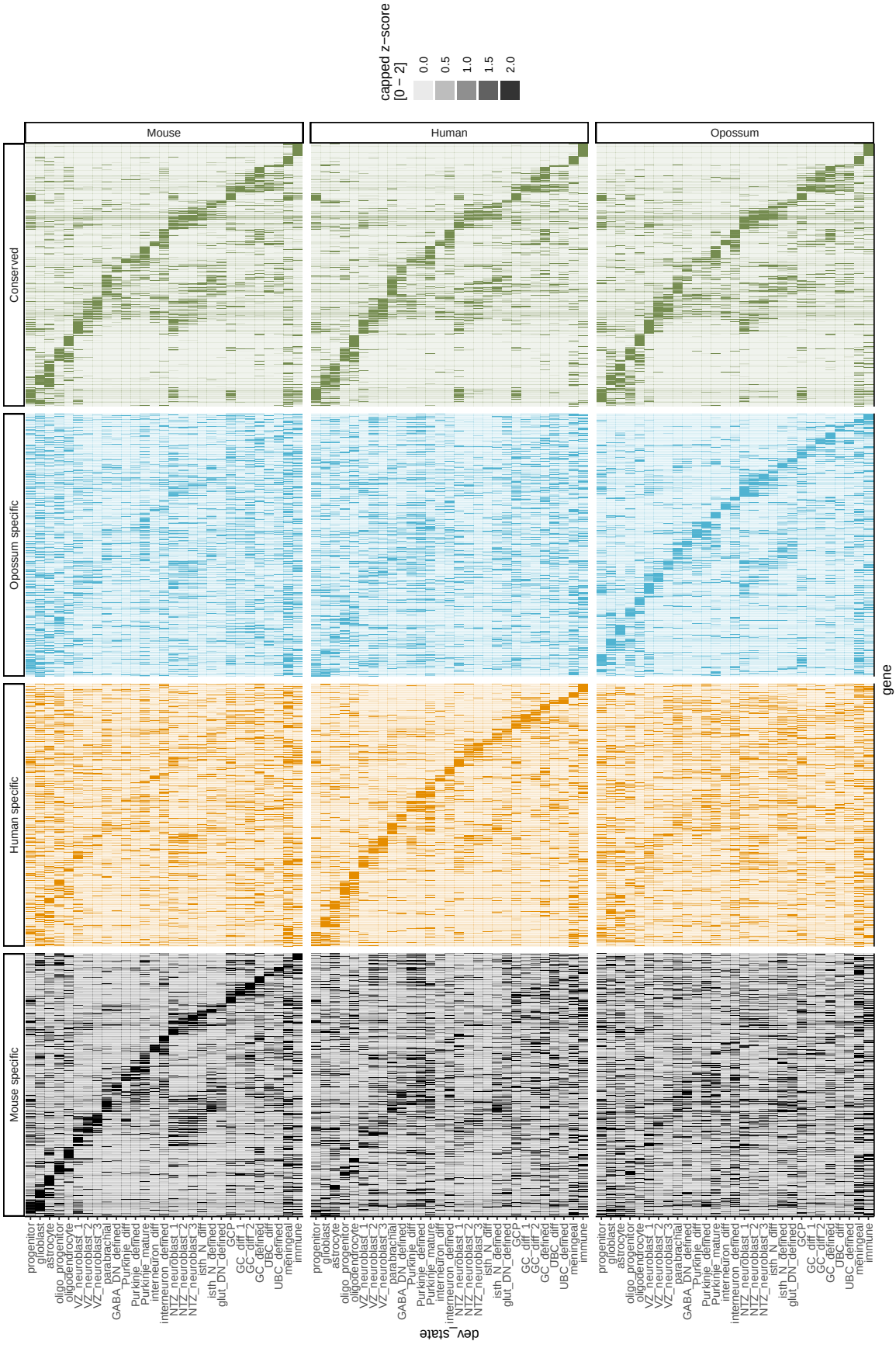


Figure 11: Overlap of cells state marker genes between species Number of occurrences of a marker gene in one or more cell states (facets). The x-axis shows the species or the group of species where the marker genes were called.

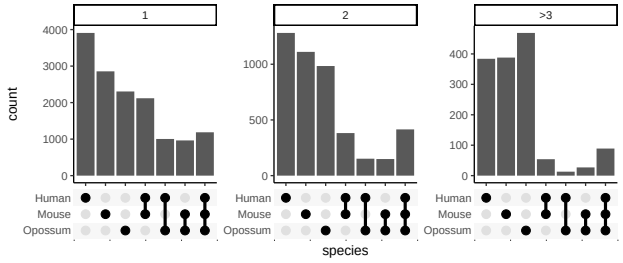


Figure 12: Overlap of conserved cell state markers between the cell states Upset plot showing conserved cell state markers identified in more than one cell state. Upset intersects are in decreasing order and capped at more than five genes.

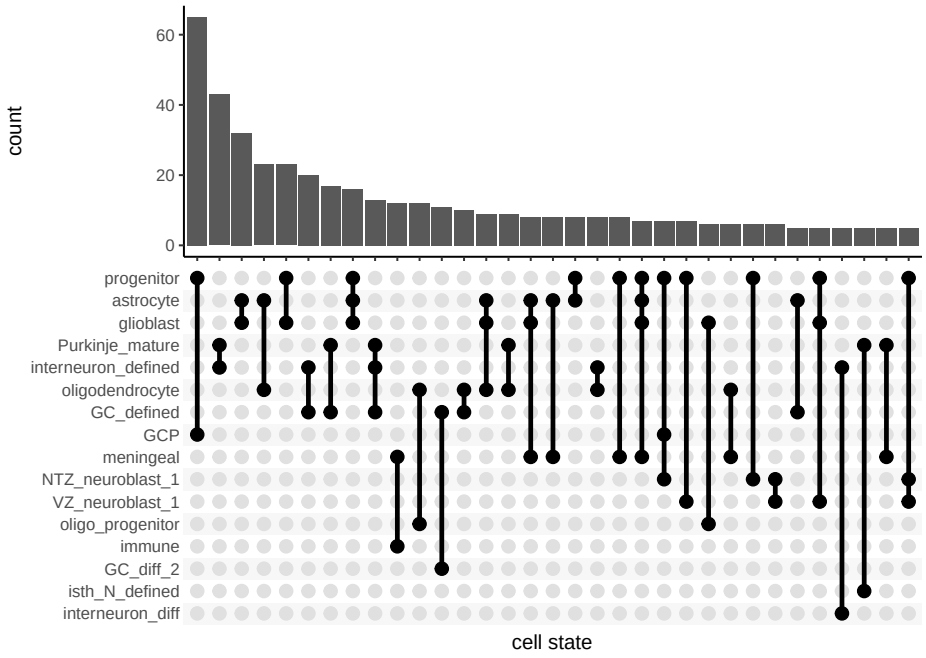


Figure 13: Conserved shared cell state marker overlap Upset plot showing all conserved cell state markers of selected neuronal cell types (x-axis). Upset intersects were ordered according to occurrence and capped at more than five genes.

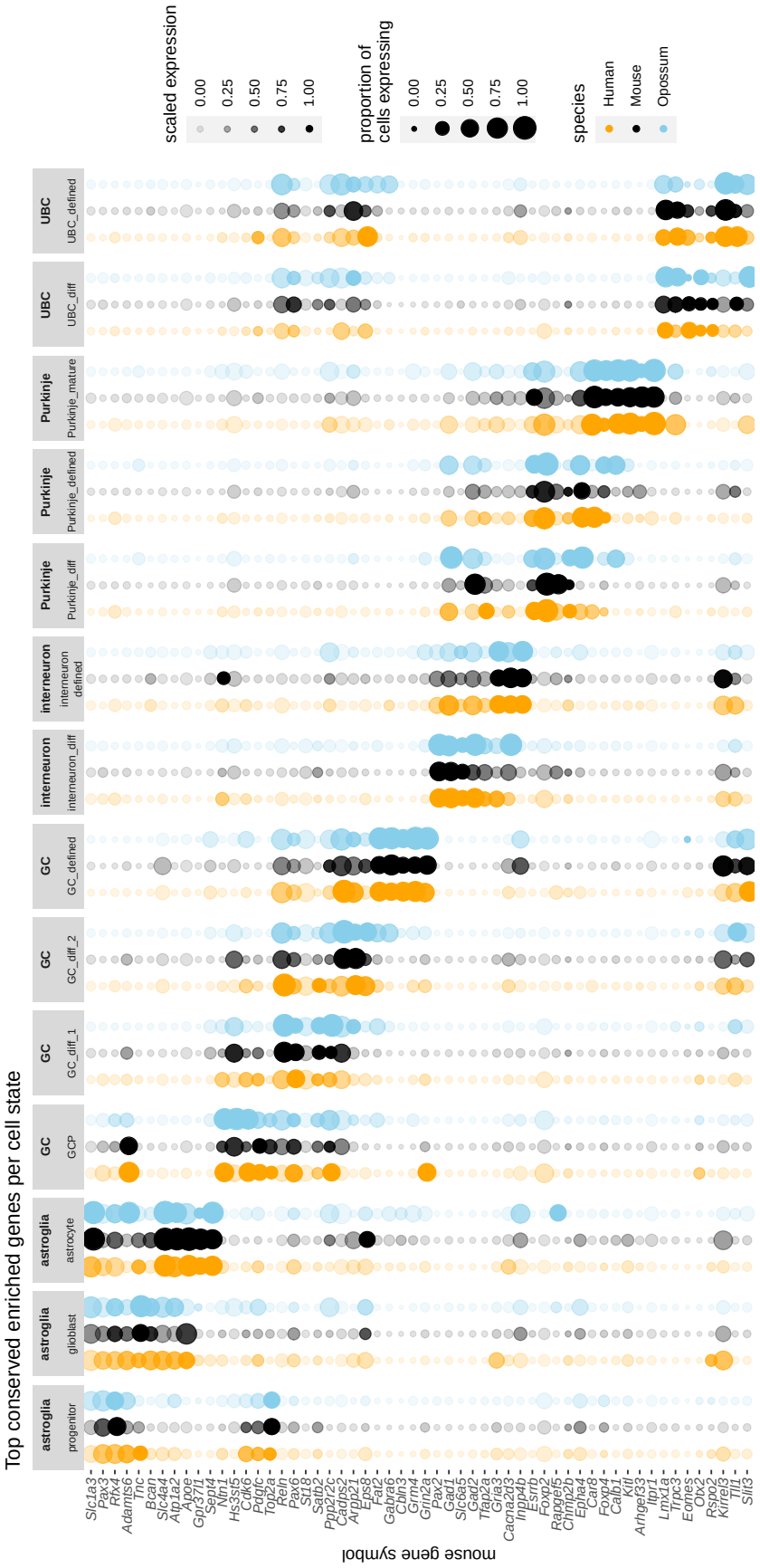


Figure 14: Top conserved marker genes for selected cell states Dotplot showing the minimum-maximum scaled expression values per species per selected cell state of the top five conserved marker genes. Size codes for the fraction of cells showing at least one UMI per group and the scaled expression value is shown as opacity of the points.

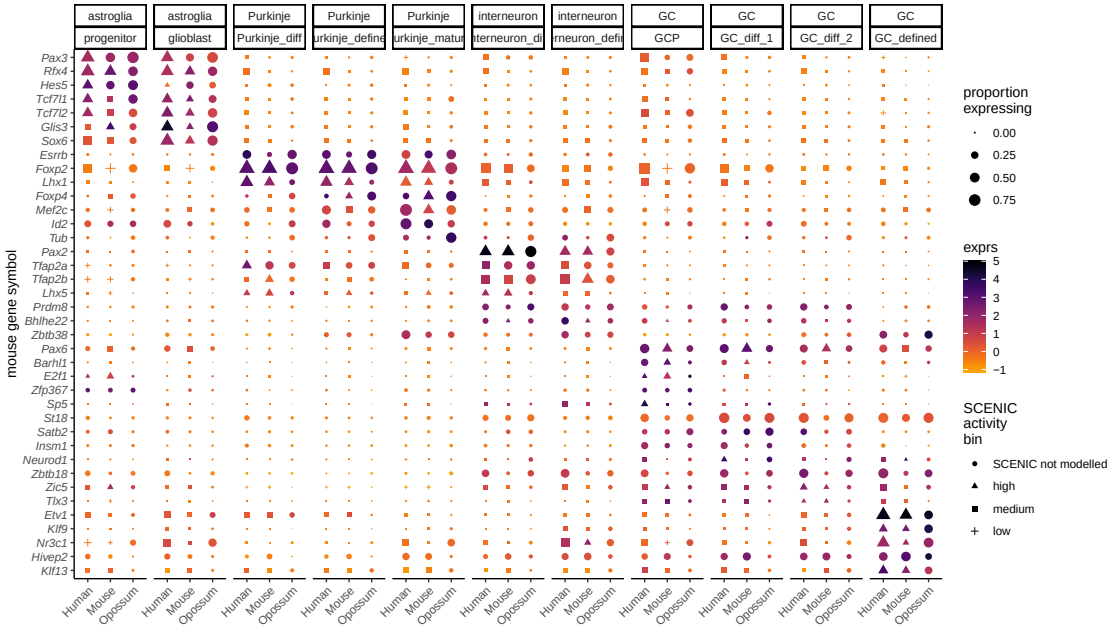


Figure 15: Conserved transcription factor expression and predicted activity Transcription factor expression was estimated on scaled expression values within the single nucleus RNA-seq dataset. Transcription factor activity was modelled for human and mouse using SCENIC. If the target network of a transcription factor could not be modelled by SCENIC, a dot is shown. Transcription factor activity (AUC) was z-scored and cut at 1 and -1. Values above 1 are assigned to high transcription factor activity, depicted as triangles. Values between 1 and -1 are assigned to medium transcription factor activity and depicted as squares. Values below -1 are assumed as low transcription factor activity and shown as crosses. Expression values were z-scored and color-coded. Top row of annotation of the facets shows the cell type label, the second row the associated cell state label.

some condensation", "meiotic spindle" and "mitotic sister chromatid segregation". Granule cell cell states exhibit an enrichment of GO terms as follows (a selection): (I) GCP: "kinetochore organization", "meiotic chromosome segregation", "positive regulation of cytokinesis". (II) GCdiff1: "neuron migration", "regulation of gene expression", "negative regulation of cell proliferation". (III) GCdiff2: "signal transduction involved in regulation of gene expression", "synaptic vesicle membrane", "axon guidance", "postsynaptic membrane" and "synapse". (IV) GCdefined: "glutamate-gated calcium ion channel activity", "locomotion", "GABA-A receptor complex" and "glutamate-gated calcium ion channel activity". Oligodendrocyte associated cell states showed first (OPC) enrichments in oligodendrocyte differentiation GO terms, such as "oligodendrocyte development" and more general in "cell maturation", followed by myelination related GO terms in mature oligodendrocytes. Mature Purkinje cells overrepresented GO terms associated with Purkinje cell maturation, "axon development" and "synaptic transmission, GABAergic".

Marker genes which are conserved and shared between cell states, are mainly overlapping in cell states which are closely related (figure 13), for example 63 genes mark both progenitors and GCPs, and 32 markers are shared between astrocytes and glioblasts. Cell states differentiating, such as VZ and NTZ neuroblasts, differentiating interneurons (interneurondiff), GCs and UBCs exhibit an enrichment of transcription factors amongst the conserved marker genes (figure S6). In total,

Table 1: Examples of known cell type marker transcription factors

Cell type / state	Transcription factor [mouse gene symbol]
progenitors	*Pax3*
astrocytes	*Hopx*
Purkinje	*Foxp2*
	Esrrb
interneurons	*Pax2*
granule cells	*Pax6*
	Etv1

there are 187 transcription factors among the conserved cell state markers. To investigate whether the apparent conserved expression of transcription factors agrees with their activity in regulating downstream effector genes, I modelled the transcription factor activity using pySCENIC [86, 87]. Using known cell type, or state, specific transcription factors, known from literature [4, 88–90], such as transcription factors listed in table 1, I compared the observed expression patterns across all three species with the SCENIC modeled activity in human and mouse (figure 15). Since at the time of writing, there was no opossum-specific *RCisTarget* database available, the marsupial was left out of the SCENIC runs. Some transcription factor activities could not be predicted in human or mouse, either due to the lack of sufficient expression, low number of identified target genes, or unknown binding motifs, which are important for the SCENIC pipeline.

Overall this analyses show that a core set of genes could be identified, which are conserved for at least 160 million years, that show cell state-specific enrichment. These genes could be assumed to be cell state defining due to their conservation. Especially, transcription factors are prime candidates for cell type identity [91]. The high overlap of SCENIC models and the detected transcription factor activity confirms the importance of the conserved transcription factors.

3.7.5 Characterization of Purkinje cell subtypes

The following section is focussed on a more detailed analysis of the Purkinje cells. More precisely, it descibes the Purkinje subtype and cell state distributions and their gene expression patterns. As mentioned previously, Purkinje cells are a major defining cell type of the cerebellum and exercise key signal integration functions.

When Dr. Mari Sepp and I studied the heterogeneity of mouse and opossum Purkinje cells, we were able to identify four groups of cells, which we assumed to represent distinct subtypes (figure 16.A). I performed marker gene enrichment for the four subtypes to investigate gene expression patterns in differentiating Purkinje cell subtypes. Four genes, ranking high in the enrichment, captured the observed diversity: *Rorb*, *Foxp1*, *Cdh9*, and *Etv1* (figure 16.D). When considering the developmental stages of subtype emergence and spatial organization of the developing cerebellum, using the Allen developing mouse brain atlas [80], spatiotemporal patterns emerged that were re-

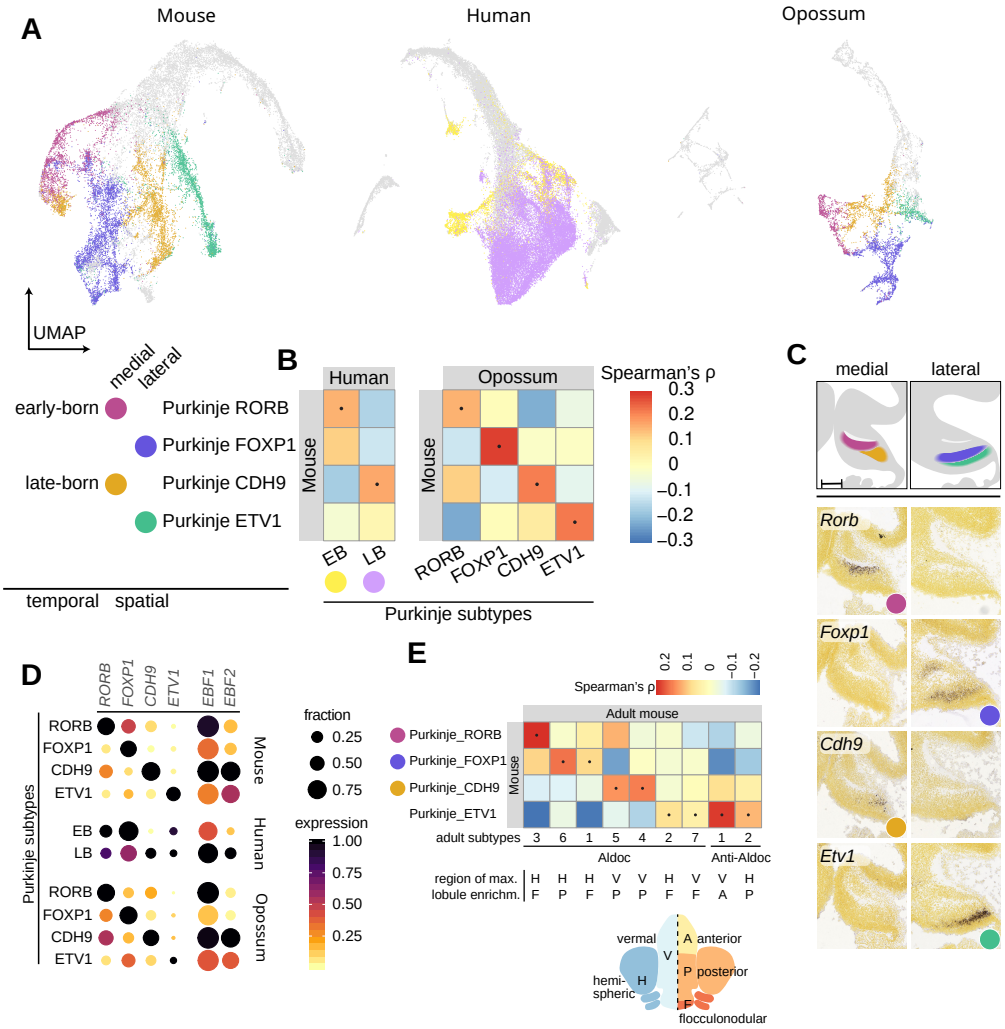


Figure 16: Purkinje cell diversity In each species Purkinje cells and ventricular zone neuroblasts assigned to the Purkinje cell lineage were separately re-integrated. **A**: UMAP embedding of the re-integration using LIGER. Identified subtypes are color-coded, nuclei which did not receive a subtype label are shown in light grey. **B**: Spearman correlation coefficients between mouse and human or mouse and opossum subtypes. Dots represent comparisons with the highest correlaton coefficient. **C**: Scheme on developing mouse cerebellum at stage E13.5 with areas of mapped location of the subtypes marked accordingly (made by Dr. Mari Sepp). Below are ISH images showing the signals for the four subtype marker genes *Rorb*, *Foxp1*, *Cdh9*, and Etv1. **D**: Dotplot of marker gene expression. Size of dots represents proportion of cells per group showing at least one UMI of the respective gene, color codes the minimum-maximum scaled expression levels of the gene. **E**: Correlation analysis between the mouse developmental Purkinje subtypes and adult subtypes identified by Kozareva et al. [9]. Below the heatmap, the adult cerebellar region where the subtype shows maximum enrichment is shown (by Dr. Mari Sepp).

flected in the expression of the aforementioned genes (figure 16.C). Early-born Purkinje cells express *Rorb* or *Foxp1*, whereas late-born cells express *Cdh9* or *Etv1*. Medially located Purkinje cell subtypes are marked by *Rorb* or *Cdh9* expression, and laterally located subtypes express *Foxp1* or *Etv1*. Furthermore the expression of transcription factors Ebf1 and Ebf2 complete the combinatorial patterning of the captured developmental Purkinje cell diversity. Medially located subtypes show higher Ebf1 expression than laterally located subtypes; late-born subtypes exhibit higher Ebf2 expression compared to early-born Purkinje subtypes. When conducting GO-term enrichment on the highly variable genes between the Purkinje subtypes, I observed an enrichment in the GO term "homophilic cell adhesion" ($p < 0.01$), a GO-term which contains many Cadherin family genes.

When studying the diversity Purkinje cells in human, only partial matches to the mouse and opossum subtypes could be found (figure 16.A,B,D): *EBF1* and *EBF2* expression mainly captures the difference between early- and late-born cells, the medial to lateral signal could not be detected.

Using orthologous gene expression of variable genes ($n = 107$), Spearman's correlation was computed (figure 16.B). Highest subtype to subtype correlation between mouse and opossum was reached for matching cell types ($\rho \approx 0.3$). The mouse to human comparison scored highest for mouse Purkinje *Rorb* and human early-born Purkinje cells, and mouse Purkinje *Cdh9* and human late-born Purkinje cells with being approximately 0.1.

The adult cerebellum exhibits a complex structural compartmentalization which is focused around Aldoc (Zebrin II) positive and negative Purkinje cells, organized in parasagittal stripes [4, 9, 62]. Kozareva et al. produced snRNA-seq data of adult mouse cerebellum and specified 9 adult Purkinje cell suptypes. [9]. I used this dataset and created pseudobulks of the adult Purkinje cell subtypes. Pseudobulks of mouse developing Purkinje cell subtypes described in this study were then compared with the adult Purkinje cell subtypes of Kozareva et al. Even though the significant gap in sampling is present between the described subtypes, possible matches between the developing and adult subtypes can be proposed based on Spearman's correlation, calculated on 337 shared highly variable genes (summarized in table 1, figure 16.E).

These results demonstrate that Purkinje cell diversity is defined both by birth date and place during early development.

3.7.6 Diversity of GABAergic interneurons

I performed similar analyses as done for Purkinje cells for the GABAergic interneurons (figure 17). After integration, clustering and marker gene enrichment analysis, five interneuron subtypes could be distinguished in mouse: early interneurons (*Zfhx4*), granule cell laye interneurons (*Rgs6*), Purkinje cell layer interneurons (*Klhl1*, *Nxph1*) and two subtypes of molecular layer interneurons (ML1 - *Sorcs3*, ML2 - *Nxph1*) (figure 17.B). The identified subtypes in the developing cerebellum could be directly matched to the subtypes found by a previous study in adult mouse cerebellum [9]. Using

Table 2: Adult mouse Purkinje groups matching to developmental Purkinje cell subtypes

Adult mouse Purkinje group	Developing Purkinje group
Aldoc$^+$ subtype 1	Purkinje *Foxp1*
Aldoc$^+$ subtype 2	Purkinje *Etv1*
Aldoc$^+$ subtype 3	Purkinje *Rorb*
Aldoc$^+$ subtype 4	Purkinje *Cdh9*
Aldoc$^+$ subtype 5	Purkinje *Cdh9*
Aldoc$^+$ subtype 6	Purkinje *Foxp1*
Aldoc$^+$ subtype 7	Purkinje *Etv1*
Aldoc$^-$ subtype 1	Purkinje *Etv1*
Aldoc$^-$ subtype 2	Purkinje *Etv1*

pseudobulk Spearman's correlation analysis ($0.3 < < 0.7$, figure 17.C). A spatiotemporal patterning could be observed: the interneuron subtypes emerge in the temporal order of (I) early interneurons, (II) granule cell layer interneuron, (III) Purkinje cell layer interneurons and (IV) molecular layer interneurons (1 and 2); and this ordering is reflected in the known spatial distribution of the subtypes in the adult cerebellum (figure 17.D).

The same subtypes could be detected in human and opossum. All aforementioned marker genes are conserved in all species and subtypes. In opossum, an additional subtype was found that expresses *MEIS2*. However, these cells mainly originate from a single individual, so further studies are needed to elucidate if these cells represent a true cerebellar cell type in the opossum.

3.7.7 Diversity of glutamatergic neurons and astroglia

Besides the aforementioned Purkinje cell and interneuron subtypes, I was able to cluster and identify, together with Dr. Mari Sepp, subtypes and cell states of other cell types. This section summarizes the characterized diversity. In total 48 subtypes could be distinguished, out of which 26 were present in all three species (table 10). The differences in the representation of subtypes between species are mostly associated with the differences in sampling.

RL/NTZ lineage shows two subtypes of glutamatergic deep nuclei neurons, which could be associated to spacial patterning. Glutamatergic deep nuclei subtype neurons (*Lmo3*) are located ventrally; the other subtype (*Lmx1a*), shows enrichment in the posterior portion of the developing cerebellum. Both subtypes reach high Spearman correlation coeffients ($\rho > 0.5$, 225 shared highly variable orthologs) when compared across species. The ventral subtype appears first, at E12.5 in mouse, the relative abundance of the posterior subtype gradually increases during development up to birth, when the strong increase in granule cells purges the abundances of glutamatergic deep nuclei cells. The other NTZ/RL-related cells were isthmic nuclei neurons. This cell type was robustly detected in all three species, with two shared and one subtype detected in mouse and opossum only. The shared subtypes are marked by *Nr4a2* or *Sst*. A subtype expressing *Slc5a7* was not detectable in human. Correlation coefficients between the species for each subtype was lower than

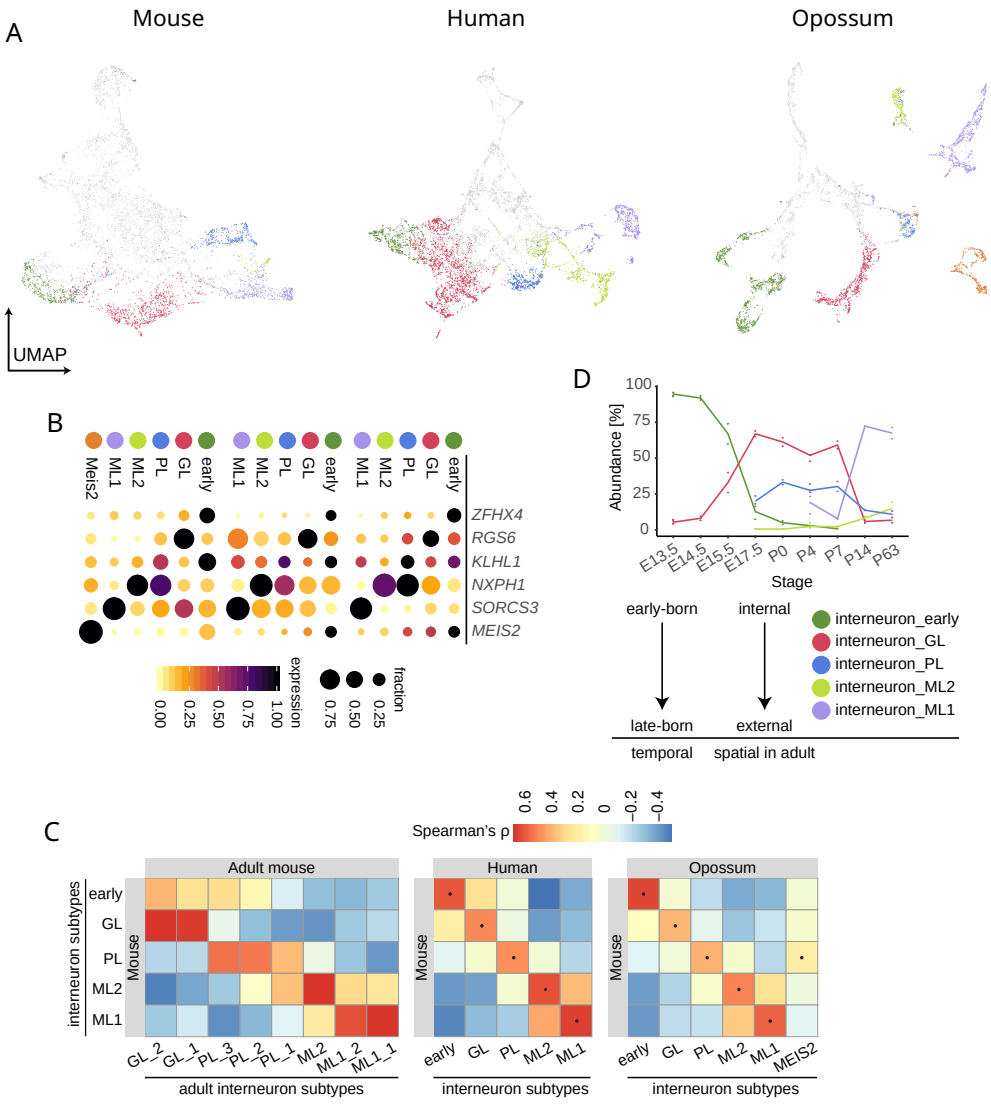


Figure 17: Interneuron diversity In each species interneurons and ventricular zone neuroblasts assigned to the interneuron lineage were separately re-integrated . **A**: UMAP embedding of LIGER guided re-integration per species. Identified subtypes are shown in different colors. Cells that were not assigned to a subtype are shown in grey. **B**: Dotplot of scaled subtype marker gene expression values (minimum-maximum scaling). Size represents the fraction of cells per group showing at least one UMI per subtype. Expression values are color-coded. **C**: Spearman correlation coefficients between pseudobulks generated per subtype. Left: Comparison of mouse subtypes with previously reported adult interneuron subtypes [9]. Center and right: Comparison between mouse and either human or opossum. Tiles with highest correlation coefficient per subtype are marked with dots. **D**: Dynamics of mouse interneuron subtype relative abundances across development. The scheme below shows relationships between temporal patterns during development and spatial distributions in adult.

for the glutamatergic deep nuclei neurons subtypes, but still clearly matchable across species when compared internally ($0 << 0.3$).

Unipolar brush cells and granule cells embed continously with each other in all three species. Dr. Mari Sepp and I were able to differentiate two UBC subtypes, one of which expresses the known UBC marker *Eomes* in addition to known UBC subtype markers such as *Trpc3*, *Grm1*, and *Calb2* [4, 9]. Furthermore, we identified a previously unknown subtype expressing *Hcrtr2* but low *Eomes* expression. Mapping *Hcrtr2* expression on the mouse developing cerebellum using the Allen Developinmg Mouse Brain Atlas, *Hcrtr2* marks a scattered group of cells in the granule cell layer. Granule cells foremost clustered into an early (*Pax6*) and late (*Synpr*) group of cells in all three studied species. In mouse and opossum, I was also able to capture a subtype expressing *Kcnip4* and *Otx2*, which was not detectable in human. Comparing the here presented mouse data to published adult mouse data [9], which contains spacial annotation, the early group does not exhibit spatial specialization. The late group shows evidences to correspond to the posterior hemisphere group and the *Kcnip4* expressing subtype might be associated to the nodulus.

Heterogeneity in the progenitor populations in the studied species was detectable, including a group of cells that might be apoptotic (low *Nckap5*, high *Bcl2l11* expression). Early progenitors reveal spatial signals: from anterior ventricular zone progenitors (*Lgr5* and *Pax5*) to posterior VZ progenitors (*Clybl* and *Cyp26b1*) and the RL progenitors (*Slit2*) . During cerebellar development, the role of the progenitor pool changes, from production of neuronal cell types to glial cells. This switch is also captured in the three species. The number of bipotent progenitors gradually increases, outnumbering purely neurogenic progenitors, during fetal and early postnatal stages (referencing here to mouse stages). Bipotent progenitors that give rise to interneurons and parenchymal astrocytes [4, 92], and gliogenic progenitors that give rise to parenchymal and Bergmann astrocytes display a decreased cell cycle score, compared to the early progenitors. Furthermore, two types of glioblast were readily detectable in the three species studied here: prospective white matter glioblasts and astroblasts. Human to mouse and opossum to mouse pseudobulk Spearman's correlation analysis showed agreement between the identified subtypes ($0.2 < p < 0.6$).

The above analyses demonstrate that the cellular diversity in the developing cerebellum is highly conserved in mammals, even at the level of subtypes.

3.8 Gene trajectories along neuronal differentiation

3.8.1 Pseudotime framework to model differentiation

Cell type differentiation is a gradual process which can not be easily modelled with clustered data. To meticulously characterize the differentiation processes, I decided to focus on the two major cerebellar neuron types: Purkinje cells and granule cells. For both cell types, due to their high

abundance during consecutive timepoints along mammalian cerebellar development, plenty of cells could be profiled and the transition from differentiating to defined cell states was captured. The UMAP embedding (figure 8.A) also reflects the process in projected transcriptomic space, where both neuronal lineages exhibit a continuum-like structure. To model Purkinje cell differentiation process, I included ventricular zone neuroblasts from the developmental time window when Purkinje cells are generated in the cerebellum (E12.5-E13.5 in mouse, 7-8 wpc in human, P4-P5 in opossum; see Methods for details).

First, the order of cells at different states of differentiation must be established. The main idea is that cells of similar differentiation state should exhibit a comparable transcriptomic profile, which separates these cells from cells at earlier or later states of the differentiation process. If enough cells were captured along the differentiation trajectory, a low dimensional embedding of the transcriptomic space should exhibit a stretched, continuum-like structure. This structure is assumed to capture and represent the signal of differentiation. After the low-dimensional embedding is established, a vector of differentiation must be fitted to the projection, in a way, that every cell is associated with a continous value, representing its progression through differentiation, often called pseudotime. I chose the classical diffusion pseudotime (DPT) algorithm as the method to call pseudotimes, given its well-established framework and straightforward interpretability,. Next, genes which show dynamic expression along the pseudotime-modelled path of differentiation, need to be identified. Correlation analysis is not sufficient to capture all the varying genes, due to the linearity assumption used by most correlation analysis methods. I tested various other approaches to identify genes that show a differentiation related linear or non-linear expression trajectory. One method that I tested, for example, leverages mutual information (MI) between pseudotime and gene expression values, but this approach had mixed results. The main issue is to set a MI threshold, which separates genes with dynamic expression from background noise [67]. This issue could be solved by using an approach that is based on the binning of the pseudotime vector und using the bins as units to call HVGs, similarly as I did for the cell type clusters, previously. This approach allowed me to call genes with dynamic expression along the pseuodotime-modelled differentiation without the assumption of linearity, and by using an explainable cut-off (based on overdispersion). Next, I assumed that there are fundamental patterns of expression along the pseudotime-modelled differentiation, hence I applied fuzzy clustering to group genes based on their expression trajectories (details in the following sections).

Since I aimed to characterize neuronal differentiation simultaneously in human, mouse, and opossum, and seeked to identify similarities and differences between the species, I needed to make sure that I compare gene expression profiles from cells that are at similar points of differentiation. First, I integrate the cells of a specific cell type from all species, and then call the pseudotime values on the integrated projection. This allowed me to skip the error-prone post hoc alignment

and to be sure that cells of similar global transcriptomic status of all three species are in the same neighborhood. To assert that this assumption holds true, I visualized the distribution of pseudotime values for the aligned developmental stages of all species (figures 19.E, 18.E).

In the final framework, cross-species integration was accomplished using the Harmony tool [76], based on a PCA, learned on a three-way one-to-one orthologous expression matrix. The root cell of the integrated embedding was chosen based on the UMAP embedding, and diffusion pseudotime was calculated. To call HVGs, I only considered genes, if a three-way one-to-one ortholog was called highly variable in all three species. Pseudobulks were created per pseudotime bin for each biological replicate. For the clustering of gene expression trajetories, each species contributed its specific ortholog trajectory, and free clustering of ortholog triplets was allowed within the given number of fuzzy clusters (n=8). To establish the order of fuzzy clusters along the pseudotime vector, hinting on the time of highest expression of a given cluster, I calculated the center of mass for each cluster and used these values for ordering the clusters (figure 19.G, 18.G). This framework was applied separately to Purkinje cells and granule cells. The results of these analyses are discussed in the following sections.

3.8.2 Purkinje cell differentiation

Purkinje cell pseudotime values distribute across developmental stages unequally. The earliest matched timepoint E12.5 shows the lowest pseudotime values (median = 0.6; figure 18.E). All following time points aggregate at the upper 10%-percentile of the pseudotime values (figure 18.D,E). This observation is shared between the species, though it is worth mentioning, that the opossum dataset lacks the group of very early Purkinje cells. The assignment of cell states reveals a consistent distribution, within all species, throughout pseudotime values (figure 18.F): early ventricular zone neuroblasts group 1 showed the lowest pseudotime values, and the value gradually increases in the row of VZ neuroblasts group 2 and 3, differentiating Purkinje cells (Purkinjediff), defined Purkinje cells (Purkinjedef), and mature Purkinje cells. This ordering is accompanied by very similar interquartile ranges for all cell states, with the exception of larger ranges for VZ neuroblasts group 2 in all species. These results are in agreement with the known transient mode of Purkinje cell emergence [2, 4], indicating that the pseudotime model accurately captures the differentiation of Purkinje cells.

The number of identified one-to-one orthologous HVGs, i.e. genes that have dynamic expression profiles during Purkinje cell differentiation, was: human = 3,651; opossum = 5,676; and mouse = 3,133. The intersection of these genes summed up to 1,846 genes, which were studied in the following analyses. Among these shared dynamic genes, there was a significant enrichment of transcription factor genes (binomial test, $p < 0.01$).

Shared dynamic genes were assigned to eight fuzzy clusters, based on their scaled expression

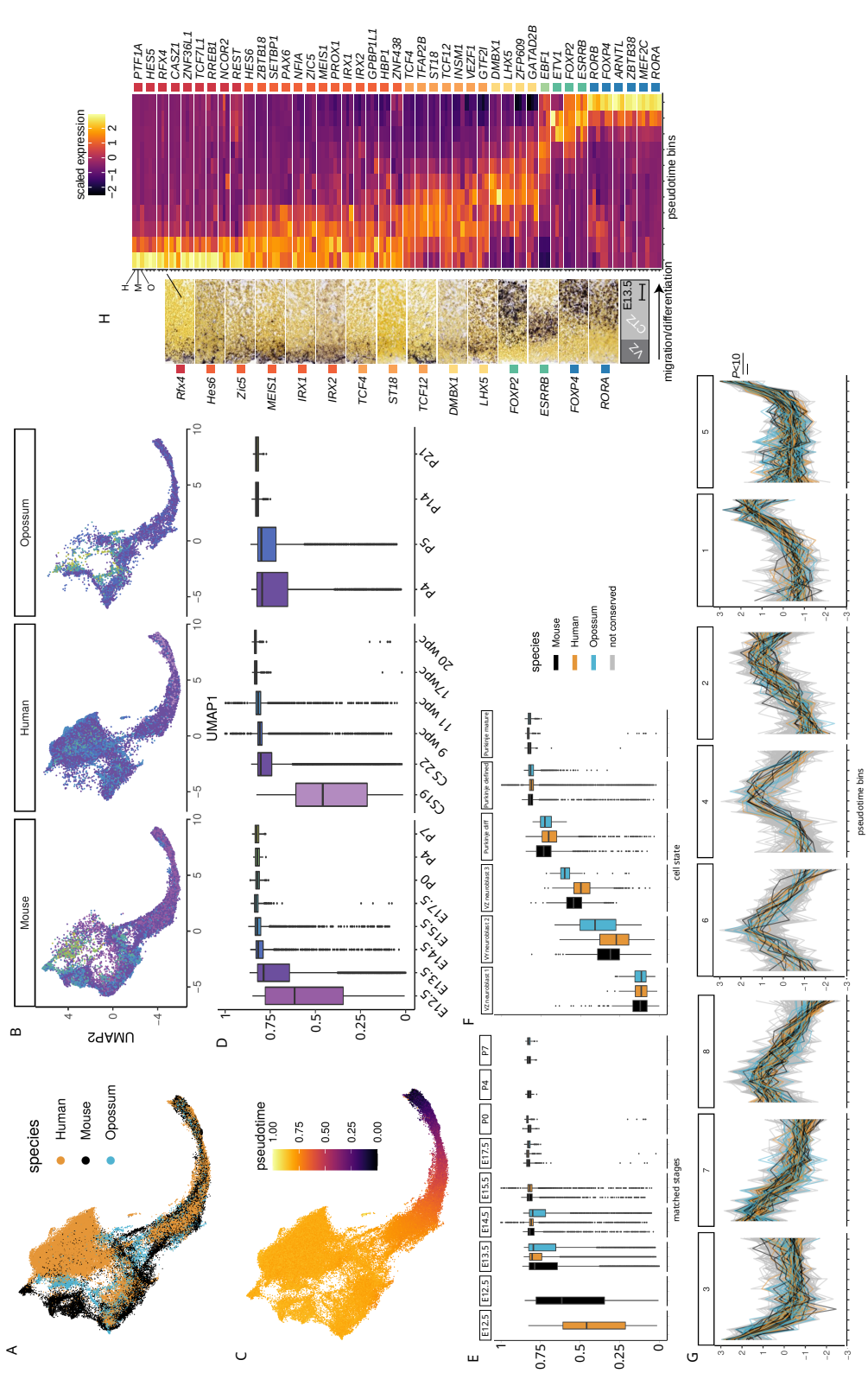


Figure 18: Purkinje cells differentiation Purkinje cells and ventricular zone neuroblasts assigned to the Purkinje cell lineage from all three were integrated using Harmony and diffusion pseudotime was detemined. **A**: UMAP embedding of the integrated data, colored according to the species. **B**: Same UMAP as in A, but split by species and colored according to matched developmental stages. **C**: UMAP embedding and colored by pseudotime values. **D**: Distribution of pseudotime values across developmental stages. **E**: Distribution of pseudotime values across matched developmental stages. **F**: Pseudotime distribution across cell states. **G**: Fuzzy clusters of gene expression trajectoryies. Colored trajectories are species assigned and belong to the conserved group. Grey trajectories belong to the intermediate or diverged groups. **H**: Heatmap of conserved transcription factor expression. Expression was z-scored and all three species are shown in triplets. On the left: ISH signals [80] of selected transcription factors at E13.5 in mouse, including overview scheme (ISH visualization and scheme done by Dr. Mari Sepp). In D-F, boxes represent the 25% to 75% percentile range, the line represents the median and the whiskers extend to maximally 1.5 time the interquantile range. Observations beyond the whiskers are shown as individual points.

profiles. These clusters exhibited varied gene expression trajectories (figure 18.G), but could be divided into three groups based on their temporal expression peaks: expression of genes in clusters 3, 7, and 8 is the highest in early differention (pseudotime bins 1-2), expression of genes in clusters 8, 6, 4, and 2 peaked transiently during differentiation, whereas genes in clusters 1 and 5 clearly showed the highest expression levels in late differentiation.

Since each orthologous gene was assigned to a cluster independently, I could evaluate the agreement between the species on a gene by gene basis. To test for differences in cluster assignments between the orthologous genes I calculated a p-value and classified each gene into one of the following groups (figure 21.A): (I) preserved - all p-values >0.5; (II) species-specific - p-value < 0.05 for of one species against the other two and p-value > 0.5 between the other two species (the gene with species-specific trajectory needed to be assigned to a different cluster than its orthologs in the other two species); (III) diverse - p-value < 0.05 for all comparisons; (IV) intermediate - if all of the previous conditions were not fulfilled; (V) not assigned - at least one ortholog in an ortholog group did not reach a maximum membership score of 0.5. I used the opossum as an evolutionary outgroup to the eutherian mammals to polarize the trajectory changes detected in human and mouse. For instance, if the orthologous gene expression trajectories were similar in mouse and opossum, but different in human, the gene was assigned as having a human-specific trajectory. Changes detected in opossum could not be polarized, i.e. it is unclear whether the change occured in the branch leading to opossum, or whether the change was established in the eutherian lineage. These genes are denoted as 'marsupial' in table 3.

Table 3: Number of genes assigned to classes in Purkinje cell differentiation

Class	n
conserved	285
human-specific	47
mouse-specific	42
Marsupial	93
diverse	58
not assigned	1321

First, I characterized the genes with preserved (i.e. strongly conserved) trajectories during Purkinje cell differentiation in human, mouse and opossum. Similar to the previous marker gene analysis (section 3.7.4), conservation was assumed to enrich for genes which are important in cell type differentiation. A total of 285 genes were classified as exhibiting preserved trajectories during Purkinje cells differentiation. I grouped these genes by trajectory cluster and preformed a GO-term enrichment analysis based on mouse genome GO term assignments. This analysis revealed cluster-specific signals and enriched GO terms that were in agreement with the temporal expression profiles of the genes in each cluster. A selection of enriched GO terms for all trajectory clusters can be

found in table 4.

Table 4: Selected enriched GO terms for conserved genes in different trajectory clusters in Purkinje cells

Fuzzy cluster	GO term	enrichment
3	neural tube development	4.9
	cilium assembly	4.1
	ruffle	3.95
	positive regulation of cell proliferation	2.47
7	cell fate determination	5.45
	smoothened signaling pathway	4.09
	RNA binding	2.23
8	translation regulator activity	6.82
	positive regulation of neuron differentiation	5.11
	regulation of transcription, DNA-templated	2.19
6	nucleosomal DNA binding	20.58
	ATP-dependent chromatin remodeling	20.58
	transcriptional repressor activity	10.29
	cell-cell signaling	10.28
4	magnesium ion binding	24.1
	receptor binding	24.1
	integral component of synaptic vesicle membrane	24.1
	synaptic vesicle membrane	18.1
2	GTPase activity	9.2
	inhibitory synapse	9.2
	myelin sheath	8.3
	dendrite	4.9
1	phospholipid binding	7.6
	glycine binding	7.6
	protein homooligomerization	7.6
	excitatory postsynaptic potential	5.7
5	GABA-A receptor activity	3.2
	glutamate receptor activity	3.2
	adult behavior	3.2
	neuron projection development	2.6
	postynaptic membrane	2.1

Next, transcription factors, which have preserved expression trajectories, were invesitgated whether their spatial expression patterns are in agreement with the temporal expression profiles. Using ISH data from the Allen Developing Mouse Brain Atlas [80], it was confirmed that the sequential expression of the transcription factors, as detected in the pseudotime analysis, is mirrored in the migrations patterns of Purkinje cells within the developing cerebellum (figure 3.H).

3.8.3 Conserved granule cell differentiation

I proceeded with the characterization of differentiation of granule cells, the other key cell type in the cerebellum, and performed similar analyses as done for the Purkinje cells.

Granule cells are produced during an extended period of time of cerebellar development [4]. For example, in mouse the earliest granule cell precursors are detectable as early as E13.5. The external granule cell layer is established which generates granule cells up to P14 [2, 4].

Dynamic genes were called in each species (human = 5,768, mouse = 2,831 and opossum = 4,782) and subsetted for three-way one-to-one orthologs: in total 1,846 shared dynamic genes were identified. After fuzzy clustering, these orthologous genes were assigned as having preserved, species-specific or diverse expression trajectories during granule cell differentiation, using the same criteria as applied in the analysis of Purkinje cells. The distribution of assignments is reported in table 19. The majority of orthologous gene groups remained as not assigned, due to their low maximum cluster membership scores. The next biggest category is formed by genes showing preserved expression trajectories.

Table 5: Number of genes assigned to classes in granule cell differentiation

Class	n
conserved	414
human-specific	136
mouse-specific	54
Marsupial	81
diverse	56
not assigned	1,163

GO enrichment analysis of the genes with preserved expression trajectories during granule cell differentiation revealed pertinent GO terms for each trajectory cluster, as summarized in table 6.

In contrast to the Purkinje cell shared dynamic genes, no transcription factor enrichment was detected among the granule cell shared dynamic genes (binomial test, $p > 0.05$). I extracted transcription factors, which exhibited preserved expression trajectories during granule cell differentiation to evaluate, whether a spatial distribution of the expression of these genes could be observed. Scaled expression of these transcription factors through granule cell differentiation is depicted in figure 19.H, next to ISH stainings from the Allen Developing Mouse Brain Atlas [80]. For the transcription factors with available ISH data, a clear correlation between the expression trajectory and granule cell migration patterns was observed. For example, *Trp53* and *E2f2* are expressed in the external granule cell layer (i.e., early in differentiation); migratory granule cells express Lhx1 and later *Pax6*; GCs reaching the internal granule cell layer express *Etv1*.

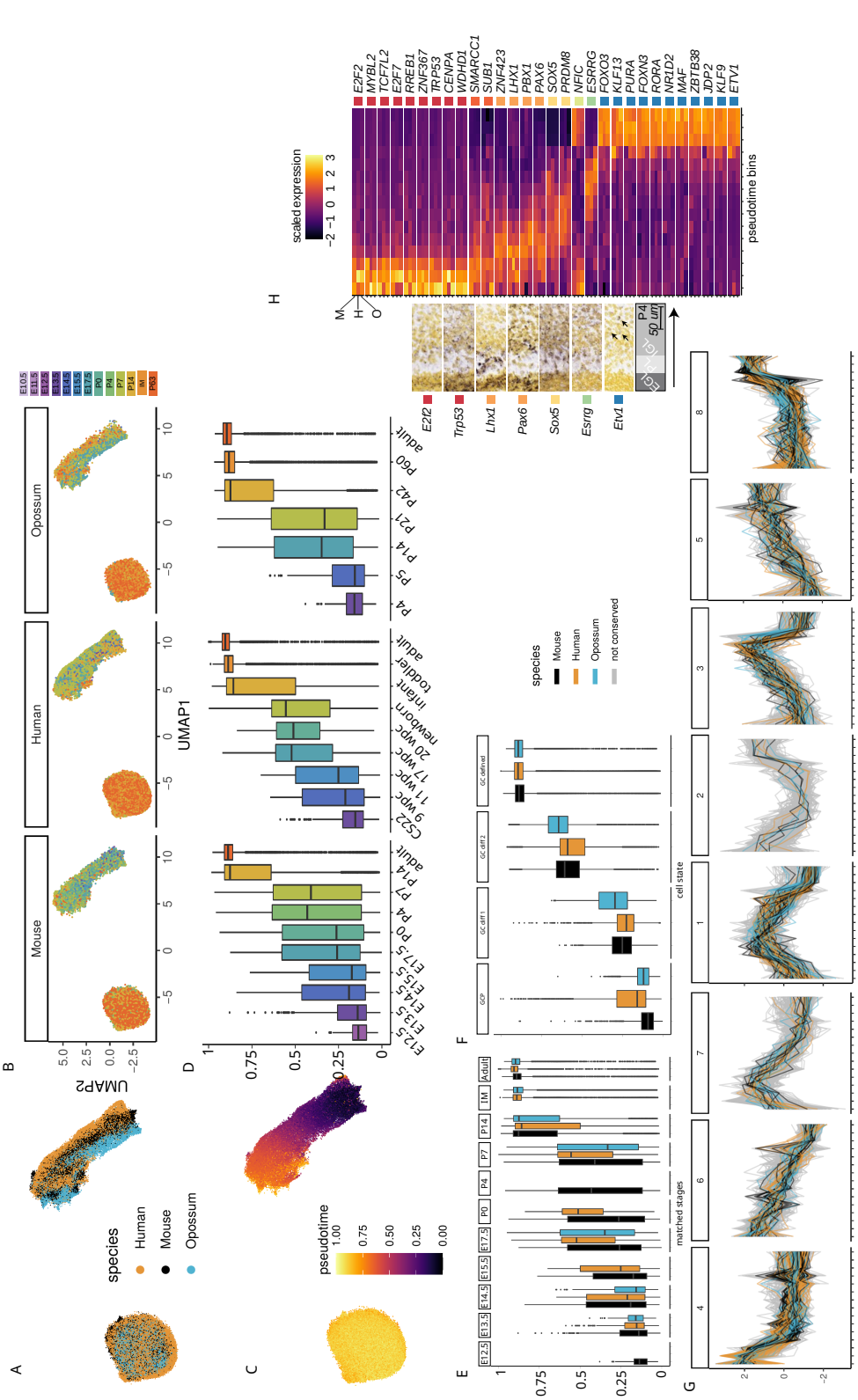


Figure 19: Granule cell differentiation Granule cells from all three species were integrated using Harmony and diffusion pseudotime was detemined. **A**: UMAP embedding of the integrated data, colored according to the species. **B**: Same UMAP as in A, but split by species and colored according to matched timepoints. **C**: UMAP embedding colored by pseudotime values. **D**: Distributions of pseudotime values across developmental stages. **E**: Distributions of pseudotime values distribution across matched developmental stages **F**: Pseudotime distributions acorss cell states. **G**: Fuzzy clusters of gene expression trajectories. Colored trajectories are species assigned and belong to the preserved group. Grey trajectories belong to the intermediate or diverged groups. **H**: Heatmap of conserved transcription factor expression. Expression was z-scored and all three species are shown in triplets. On the left: ISH signals [80] of selected transcription factors at P4 in mouse, including overview scheme (ISH visualization and scheme done by Dr. Mari Sepp). In D-F, boxes represent the 25% to 75% percentile range, the line represents the median and the whiskers extend to maximally 1.5 time the interquantile range. Observations beyond the whiskers are shown as individual points.

Table 6: Selected enriched GO terms for conserved genes in different trajectory clusters in granule cells

Fuzzy cluster	GO term	enrichment
4	mitotic sister chromatid segregation	2.5
	DNA replication checkpoint	2.5
	G2/M transition of mitotic cell cycle	2.5
	cell division	2.3
	DNA replication	2.3
	cell cycle	2.2
6	structural constituent of ribosome	21.5
	translation	13.4
	mRNA binding	10.8
	RNA binding	8.6
	regulation of translation	7.2
7	anterior/posterior pattern specification	26.1
	dorsal/ventral pattern formation	23.2
	negative regulation of neuron differentiation	23.2
	regulation of gene expression	17.4
	smoothened signaling pathway	11.6
1	positive regulation of filopodium assembly	11.4
	cell fate commitment	7.5
	dendrite morphogenesis	7.5
	positive regulation of cell migration	5.6
	cell migration	4.5
	neuronal cell body	2.4
2	heterophilic cell-cell adhesion via plasma membrane	27.7
	cell-cell adherens junction	15.8
	apoptotic process	5.8
3	membrane depolarization during action potential	14.3
	voltage-gated sodium channel complex	10.7
	neuronal action potential	10.7
	cell-cell adhesion	8.6
	learning or memory	7.1
	synaptic transmission, glutamatergic	7.1
5	memory	11.4
	vesicle	7.6
	cytoplasmic microtubule	7.6
	neurotransmitter secretion	7.6
	axon extension	7.6
	SNARE binding	5.7
	adult walking behavior	5.7
	axolemma	5.7
8	sodium:potassium-exchanging ATPase activity	3
	potassium ion import	3
	neuronal signal transduction	3
	synaptic vesicle membrane	3
	axonal growth cone	3

3.8.4 Functional relevance of the dynamic genes

Dr. Mari Sepp and I asked, whether the identified dynamic genes are enriched for genes that are essential for normal development. To address this question, I grouped genes either in cross-species shared dynamic or not dynamic groups. Furthermore, I assigned each gene into one of three bins, depending on whether and in which cell type a gene is pseudotemporal dynamic (none, Purkinje, GC, both). About 57% of all dynamic genes are shared between the two cell types. This group of genes, and Purkinje dynamic genes show signs of transcription factor enrichment: Purkinje dynamic genes exhibited an empirical enrichment of 1.5 ($p < 0.001$) and shared dynamic genes showed an enrichment of 1.2 ($p = 0.078$) of transcription factor coding genes.

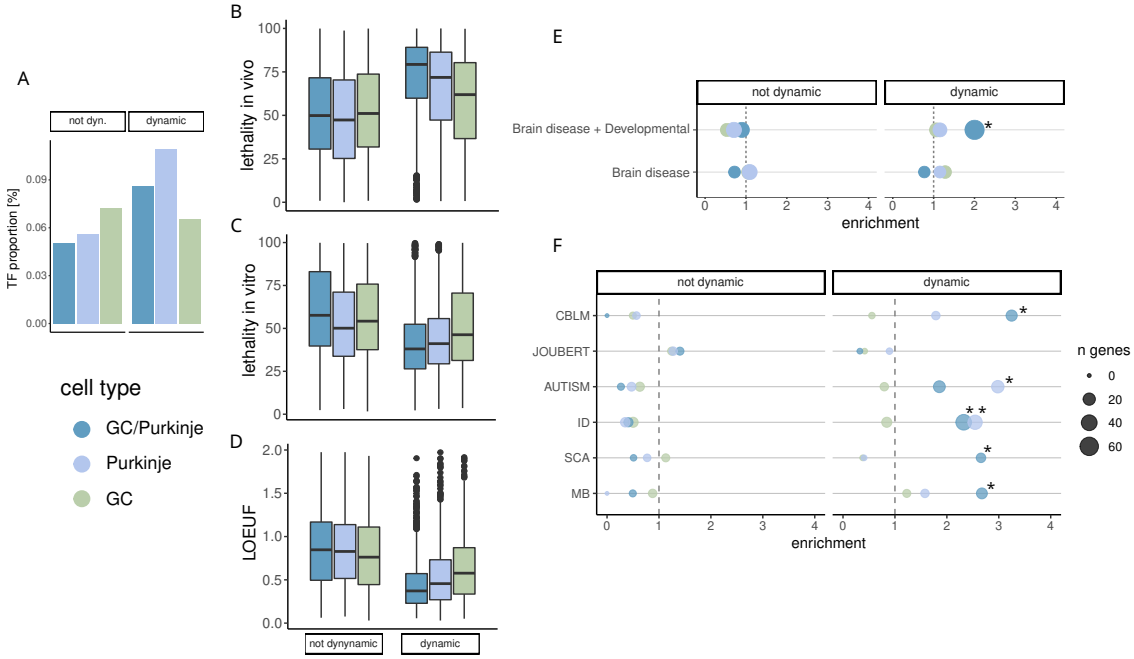


Figure 20: Functional significance of dynamic genes **A**: proportion of transcription factors among variable and non-variable genes during either Purkinje cell, granule cell or both cell types differentiation. **B**: lethality *in vivo* score for either dynamic or non dynamic genes. **C**: lethality *in vitro* score for either dynamic or non dynamic genes. **D**: LOEUF scores (loss-of-function observed/expected upper bound fraction, lower scores = higher constraint). **E**: enrichment analysis within dynamic and non-dynamic genes for brain disease associated genes either with developmental influence or without. Size of dots represents the number of associated genes, the x-axis shows the observed enrichment. **F**: enrichment analysis for cerebellum disease associated genes. CBLM = cerebellar malformations, JOUBERT = Joubert syndrome, AUTISM = autism spectrum disorder, ID = intellectual disorder, SCA = spinocerebellar ataxia, MB = medulloblastoma. Significant enrichments ($p < 0.1$) are indicated with a star.

I used an *in vivo* essentiality score that measures tolerance to heterozygous inactivation in the human population, as well as an *in vitro* essentiality score, which was determined by viability assays in human cell lines [6]. Both metrics can be used to evaluate the level of functional constraints. Compared to the non-dynamic genes, dynamic genes have a clearly higher *in vivo* gene essentiality score, whereas the in vitro essentiality scores were slightly lower (permutation test ($n = 10,000$), H_1 = 'greater',$\alpha = 0.05$). Genes that are dynamic both in Purkinje and granule cells showed

significantly higher intolerance to heterozygous inactivation in the human population, compared to genes, which were called in only one of the two cell types. Dynamic genes in Purkinje cells showed higher *in vivo* intolerance, compared to granule cell-associated dynamic genes (permutation test ($n = 10,000$), $\alpha =' greater'$, $p < 0.01$). Additionally, we used the recently described LOEUF scores (loss-of-function observed/expected upper bound fraction) [93], the most up-to-date measure of *in vivo* essentiality provided by the Genome Aggregation Database (gnomAD). Again, dynamic genes show stronger constraint than non-dynamic genes (permutation test ($n = 10,000$), $H_1 =' lower'$, $p < 0.01$), and among the dynamic genes, genes which are shared between Purkinje and granule cells exhibit the strongest constraint (permutation test ($n = 10,000$), $\alpha < 0.01$, $p < 0.01$) (figure 20).

In light of these results, I next studied possible disease associations of the dynamic genes. I obtained as list of inherited disease and gene associations from the Human Gene Mutation Database (HGDM, PRO 17.1) [94]. I subsetted this list to disease driver genes, utilizing frameworks established in our lab [66]. Furthermore, I defined 'Brain disease' genes as the genes associated to disease types 'Nervous system' and 'Psychiatric', according to the Unified Medical Language System (UMLS). Due to the developmental nature of the data in this thesis, I additionally grouped the filtered genes whether they are also associated with the high level disease type 'Developmental'. This led to the following sets of 'Brain disease' genes: developmental ($n = 373$) and other ($n = 200$). A binomial test was applied to investigate whether any dynamic or non-dynamic cell type-associated gene group is enriched in any of the described disease gene groups. Significant results ($\alpha = 0.1$) were obtained in only one group: developmental brain disease associated genes are enriched among the shared dynamic genes with an empirical enrichment of approximately two (figure 20.E).

Next, I focussed on the diseases that are directly associated with the functions of the cerebellum. I downloaded a curated list of neurodevelopmental and adult-onset neurodegenerative disorders linked genes from Aldinger et al. [95]. Genes linked to spinocerebellar ataxia, medulloblastoma or cerebellar malformations, including Dandy-Walker syndrome and cerebellar hypoplasia, showed enrichment in dynamic genes shared between Purkinje and granule cells (table 7, figure 20). Additionally, genes dynamic in Purkinje cell differentiation showed enrichment of high-confidence risk genes in autism spectrum disorder and intellectual disability.

Taken together this shows that genes, playing a role in granule and Purkinje cell differentiation, exhibit increased functional constraint compared to non-regulated genes. Furthermore developmental associated brain diseases show enrichment among both cell types. Zooming into specific syndromes and diseases, often genes associated with differentiation of both cell types are affected, indicating that the studied diseases are interfering not only with one but presumably multiple cell types in the cerebellum.

Table 7: Disease associations for genes with dynamic expression during Purkinje and granule cell differentiation

Disease	Group	Empirical enrichment	p	FDR
high confidence risk genes of autism spectrum disorder	GC	0.795	0.765	1.000
	Purkinje	2.981	0.000	0.000
	GC/Purkinje	1.856	0.010	0.054
cerebellar malformations	GC	0.557	0.874	1.000
	Purkinje	1.789	0.123	0.555
	GC/Purkinje	3.248	0.000	0.001
intellectual disorder	GC	0.843	0.755	1.000
	Purkinje	2.544	0.000	0.000
	GC/Purkinje	2.324	0.000	0.000
Joubert	GC	0.418	0.909	1.000
	Purkinje	0.894	0.655	1.000
	GC/Purkinje	0.325	0.954	1.000
medulloblastoma	GC	1.228	0.385	1.000
	Purkinje	1.578	0.184	0.737
	GC/Purkinje	2.675	0.001	0.007
spinocerebellar ataxia	GC	0.380	0.929	1.000
	Purkinje	0.407	0.915	1.000
	GC/Purkinje	2.657	0.008	0.048

3.8.5 Comparisons of functional constraints of genes with preserved or diverged expression trajectories

As described in the previous sections, besides genes with preserved trajectories during Purkinje and granule cell differentiation, I also identified orthologous genes that had diverged trajectory patterns in at least one of the studied species. To investigate the functional constraints of the genes with different level of trajectory similarities between the species, I grouped the genes with either species-specific or diverse trajectories as 'diverged' genes ($\mu = 17\%$), and compared these to the genes with preserved trajectories ($\mu = 12\%$) and with genes in the intermediate category ($\mu = 60\%$). To increase the resolution of this analysis, the genes of the three groups of conservation were split by the assigned fuzzy cluster in human, into groups of gene with the highest expression in early, middle or late differentiation (as determined by center of mass per cluster in human) (figure 21.D-F).

First, I investigated the *in vivo* functional constraints using two measures – the LOEF scores [93] and the *in vivo* intolerance scores from [6]. Genes with preserved trajectories in granule cells and expression peak in mid differentiation exhibited significantly lower LOEF scores (indicating stronger *in vivo* functional contraints) compared to diverged and intermediate genes (permutation test ($n = 10,000$), $p = 0.047$, $p = 0.089$, $\alpha = 0.1$)(figure 21.D). Considering *in vivo* intolerance scores, I only observed significantly higher (indicating stronger *in vivo* functional contraints) in

early expressed intermediate genes, compared to preserved genes in granule cells ($p=0.001) (figure 21.E). In Purkinje cells, significant decreases in LOEF scores were identified for late expressed preserved genes, compared to intermediate and diverged genes (permutation test ($n = 10,000$), $p = 0.02$, $p = 0.052$, $\alpha = 0.1$, figure 21.D). Consistently, mid and late expressed preserved genes showed significant increases in *in vivo* intolerance scores, compared to the respective intermediate and diverged genes ($p < 0.05$,figure 21.E).

Next, I investigated the *in vitro* functional constraints [6]. Human *in vitro* intolerance scores were significantly increased (indicating stronger functional contraints) for early expressed preserved genes, compared to intermediate ($p < 0.001$) and diverged ($p < 0.001$) genes in granule cell differentiation. In Purkinje cell differentiation, the early an mid expressed preserved genes showed higher in vitro functional constraints than the diverged genes ($p < 0.001$ figure 21.F).

The following comparisons showed significant differences: in early expressed preserved were increased compared to diverged genes ($p = 0.016$) and intermediate higher than diverged genes ($p = 0.0014$); For genes peaking in the middle of differentiation, preserved were higher than intermediate genes ($p = 0.004$) and preserved vs. diverged genes ($p < 0.001$, figure 21.F).

I also tested, whether any of the defined groups enrich in cancer associated genes, but none of the analyses revealed significant enrichments ($\alpha < 0.05$).

Taken together, genes with preserved expression trajectories during Purkinje or granule cell differentiation show higher levels of functional constraints than the diverged genes. However, the mode of constraint differs between the genes with different temporal expression patterns. In general, earlier expressed preserved genes are required for cellular viability (*in vitro* constraints), later expressed preserved genes are intolerant to heterozygous inactivation in humans (*in vivo* constraints).

3.8.6 Characterization of the diverged gene expression trajectories

After obtaining the global view on three-way orthologs with conserved and diverged trajectories, I zoomed in on genes with species-specific trajectory changes in granule and Purkinje cells. I used the opossum as an evolutionary outgroup to the eutherian mammals to polarize the expression trajectory changes detected in human and mouse. Differences detected in opossum could not be pinpointed to whether the change occurred in the branch leading to opossum, or whether the change was established in the eutherian lineage.

A summary of the number of orthologs called either human-specific, mouse-specific, marsupial, or diverse (no resemblance of trajectory between in any of the species) is depicted in figure 21.A. In granule cells, a total of 136 genes exhibited human-specific trajectory changes. Statistical analysis (binomial test) revealed that this is a clear enrichment, compared to the other groups ($p < 0.001$). No other group exhibited significant overrepresentation ($\alpha = 0.05$). In Purkinje cells, the numbers of human and mouse changes were comparable (47 and 42). The number of genes with species-specific

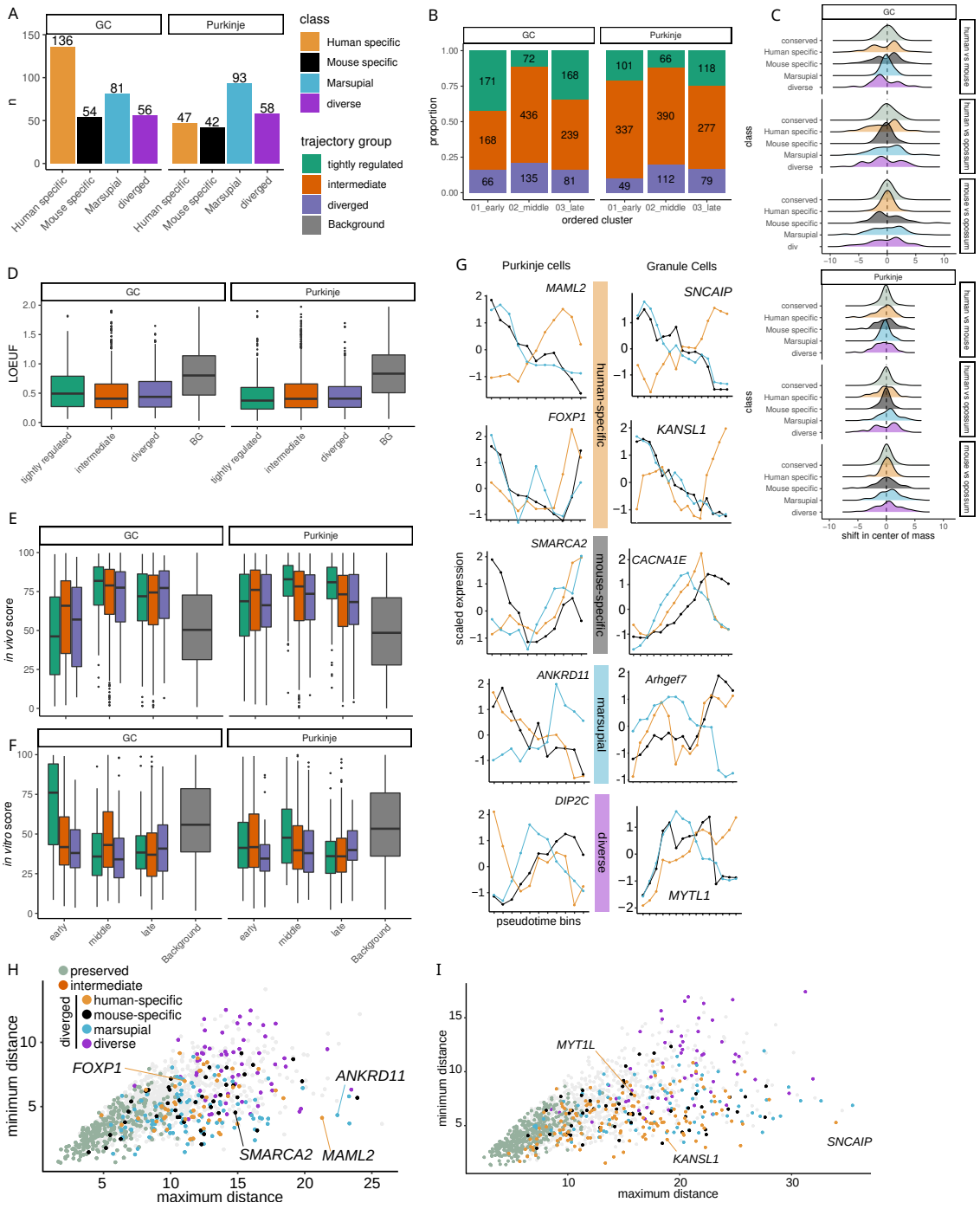


Figure 21: species-specific trajectories in Purkinje and granule cells A: number of classified trajectories per species-specific and totally diverged (no direct resemblance between any species) for Purkinje and granule cell differentiation. **B**: groups of gene expression peaks characterized by either "tightly regulated" (conserved trajectories), "diverged" (species-specific or totally different) and "intermediate" (other variable genes). Total numbers are printed and relative proportion plotted. **C**: per ortholog triplet the pairwise difference in center of mass was determined to quantify the directionality of possible changes (e.g. early to late, or late to early). Quantification was done for Purkinje and granule cell differentiation. **D**: LOEUF scores per trajectory group for both cell types (lower values = higher constraint). BG = non variable genes detectable in all species per cell type. **E**: *in vivo* lethality score per trajectory group and time of peak for human ortholog. **F**: *in vitro* lethality score per trajectory group and time of peak for human ortholog. **G**: examples of genes per trajectory group in Purkinje and granule cells. **H**: minimum vs maximum dynamic timewarp distance observed per ortholog triplet in Purkinje cells. **I**: minimum vs maximum dynamic timewarp distance observed per ortholog triplet in granul e cells. Examples are highlighted and trajectory groups color-coded.

trajectories in both Purkinje and granule cells was low (1 - 4).

Next, I investigated whether the identified changes in trajectories show biases in directionality, i.e. if the expression shifts systematically towards earlier or later differentiation. For each gene, I quantified the change in the center of mass values between the species. As expected, these values were distributed around 0 for the genes with preserved trajectories. Genes with human-specific, mouse-specific, marsupial or diverse trajectories showed distributions that deviated further away from 0, in the expected pairwise species comparisons. Importantly, no clear tendency for shift directionality could be detected by visual investigation in either of the cell types (figure 21.C).

Till this point, the classification of the orthologs was done on a categorical scale. To provide a quantitative measure of the change, I estimated dynamic timewarp distances (DTW) between all orthologous groups of genes. I calculated the DTW distances pairwise between the orthologs from the three species, and determined the maximum (D_{max}) and minimum (D_{min}) distance observed. I observed the following patterns on the DTW distance plane: genes with preserved trajectories show low minimum and maximum; genes with high D_{max} and low D_{min} diverged in one of the species; and genes with high D_{max} and high D_{min} are indiciations for a diverse trajectory pattern (different trajcetories in each species)(figure 21.H/I). Among the genes with human-specific changes in granule cells, *SNCAIP* showed the lowest for its order of magnitude, thus representing a gene with the most pronounced human-specific trajectory change in granule cells (figure 21.G). For comparison, *KANSL1* and *MYTL1* show human-specific trajectory changes that are of medium or low degree, respectively (figures 21.H and G). Similarly, in Purkinje cells, *MAML2* displays a the most pronounced human-specific trajectory change, whereas *FOXP1* is an example of genes with a mild change. Further examples of genes with trajectory changes are shown in figure 21.G

Taken together, these analyses revealed many genes that have evolved a new expression trajectory during Purkinje or granule cell differentiation.

3.9 Fundamental changes in gene expression - gains and losses

In this section I summarise the work that focuses on fundamental differences in cell type-specific gene expression profiles in the cerebella of human, mouse and opossum. Previous work from our laboratory used bulk RNAseq data covering the development of the cerebellum from different mammalian species, and identified many radical differences in developmental gene expression trajectories between the species [7]. One hypothesis raising from this work, is that the trajectory differences at the whole tissue level (bulk) might reflect events of cell type-specific expression gains and/or losses. Therefore I asked, whether there are genes that have gained or lost expression in at least one of the cerebellar cell types in the three mammalian species studied. The main challenge with this analysis was to identify a robust framework to call presence or absence of gene expression in the snRNA-seq datasets with high confidence.

3.9.1 Framework to call presence or absence of expression

In these analyses, I decided to use the exonic counts to avoid biases in detectability caused by the differences in the number of intronic adenine mononuceotide repeats between the species [96]. Also, I excluded Y-coded genes. I collapsed all cells of a given cell type and biological replicate into pseudobulks. To my knowledge, there is no framework for presence and absence calling using snRNA-seq data. Most previously published work focuses on considerable expression changes, only [97]. Therefore, I designed a custom algorithm to classify the expressed 12,756 three-way orthologs into groups, depending on their expression pattern comparison.

First, pseudobulk gene expression counts were normalized as counts per million (CPM). Next, for each cell type, a gene's maximum expression (M) across all samples from all developmental stages was determined. The maximum expression values were used to classify genes as expressed (present) or not expressed (absent) in the cell type during the development in each species. Furthermore, a minimal expression per gene to be considered present was arbitrarily set to 50 CPM.

For the comparison, the pairwise quotient of expression (Q) for each cell type between the species was calculated by dividing the maximum expression in the species of focus by M of the other two species. Additionally, to be called as expressed (present) I required that a gene's M in a cell type is at least 30% of its maximum expression across all cell types of this species (P)[1]. The rules for classification are summarized in equation 2.

$$C_{ij} = \begin{cases} n(M_j > 50) \geq 2 \wedge Q_{jk} > 5 \wedge P_j > 0.3 \wedge P_k < 0.5 & specific_j \\ n(M_j > 50) = 0 \wedge n(M_k) \geq 2 \wedge P_j < 0.3 \wedge P_k > 0.3 & lost_j \\ n(M_j > 50) \geq \wedge n(M_k > 50) \geq 2 \wedge P_j > 0.3 \wedge P_k > 0.3 & conserved \\ else & ambiguous \end{cases}, \quad (2)$$

where gene i is classified for cell type j and tested against both other species k and n the number of logical evaluations to "true".

Since technical dropouts could not be ruled out, I added another layer of filtering: if a gene was not considered to be reliably expressed (<50 CPM) in any of the cell types in the snRNA-seq dataset, I checked its expression in the bulk RNA-seq data covering the development of the cerebellum [7]. If the maximum expression levels of the gene in the bulk data were above 5 CPM, I assumed a technical issue with this gene's detectability in snRNA-seq data, and therefore excluded it from all further analyses. In total 3,271 genes of 12,756 three-way orthologs, passing the initial filters, were classified as not-resolvable due to technical limitations. For each cell type, the remaining genes were classified as follows: genes which were confidently expressed in all three species were classified as conserved among therian species. Genes which were called expressed in human but not in mouse,

[1]This filtering step was suggested by Dr. Mari Sepp and implemented by me.

or vice versa, were classified using the opossum as an outgroup species as gained or lost in human or mouse. Similar to the trajectory analysis, marsupial specific expression differences could not be assigned to a lineage (eutherian or marsupial), therefore, opossum-specific expression changes were labelled as either "Marsupial expressed" (only detected in opossum), or "Eutherian expressed" (not detected in opossum, but present in human and mouse). The genes that did not meet our stringent criteria to be called present (expressed) or absent in a cell type in at least one of the three species, were classified as "ambiguous". The application and validation of this framework is described in the following section.

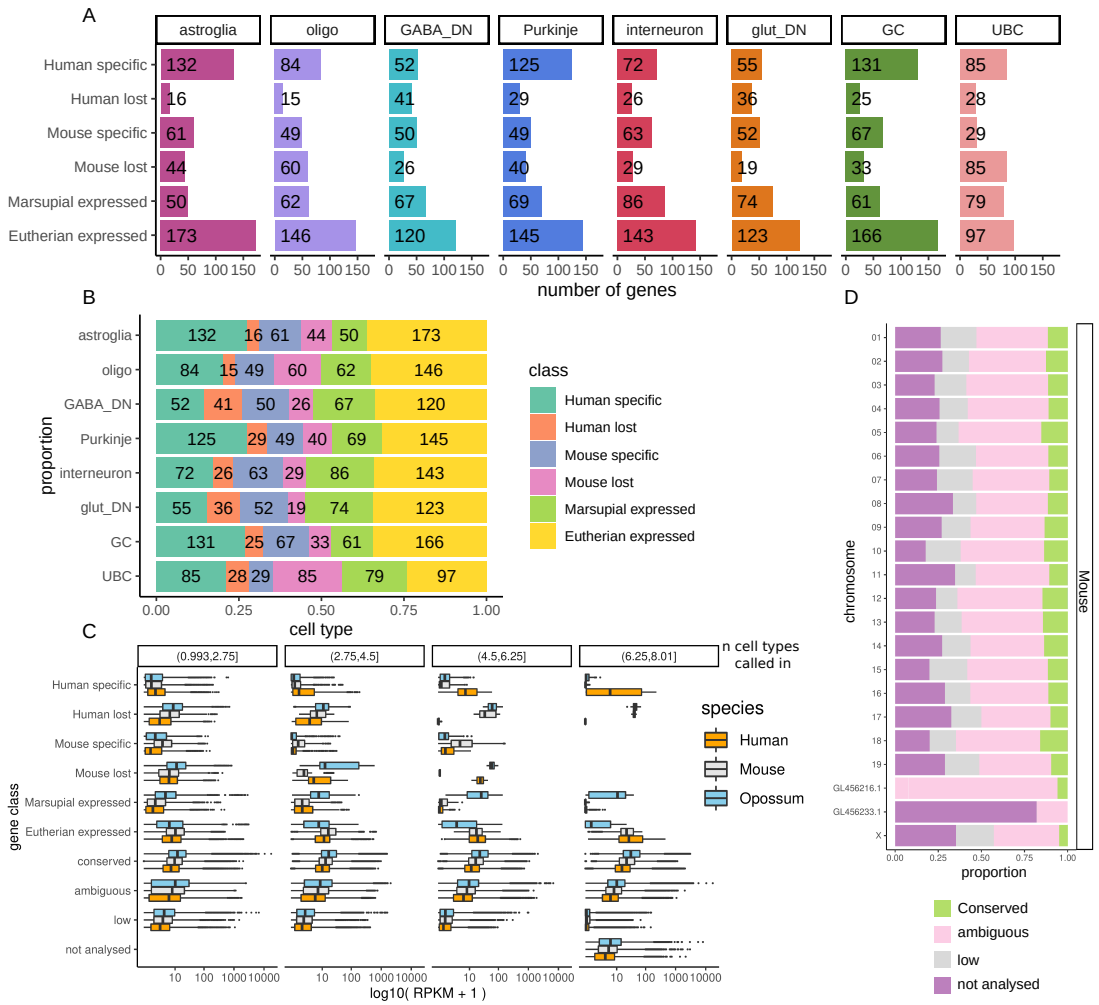


Figure 22: Gain / loss classification statistics **A**: number genes per class and studied cell type. **B**: number of genes per class and plotted according to proportion per cell type. **C**: classification quality control by using bulk RNA-seq data. The number of cell types a gene was called in a specific class was quantified and faceted. Maximum expression in bulk data was determined for all available cerebellum samples and plotted per class. **D**: classification results per mouse gene. Proportion per class per mouse chromosome was determined to study chromosome biases.

Dr. Mari Sepp and I decided to focus on the main neuronal and glial cell types in the cerebellum, which were also readily detectable in all three species: astroglia, GABAergic deep nuclei neurons, glutamatergic deep nuclei neurons, granule cells, interneurons, oligodendrocytes, Purkinje cells, and unipolar brush cells. The results of this analysis are summarised in figure 22. In all analysed cell

types, I could classify genes in all defined classes (figure 22.A). In all cell types, the vast majority of genes were classified as 'ambiguous' ($\mu \approx 4,000$), 'not analysed' ($\mu \approx 3,000$) or 'conserved' ($\mu \approx 2,300$, not shown in the figure). The number of genes in the other classes varied between 15 and 166. Across the studied cell types, the number of 'Eutherian expressed' genes and 'human gained' genes is the highest. In general, the smallest group is the class of 'human lost' genes (figure 22.A/B). Overall, the numbers of genes in different classes are comparable between cell types (figure 22.B).

To validate the framework used to call presence or absence expression in snRNA-seq data, I studied the maximum expression levels of the classified genes, grouped by the number of affected cell types, in the bulk RNA-seq data of cerebellum development [7] (figure 22.C). This analysis showed, for example, that genes which were classified as 'human gained' in more than six cell types, show a higher bulk RNA-seq signal in human than in mouse and opossum. In contrast, genes of the class "human lost" exhibited a very low signal in human compared to the other two species. 'Conserved' genes show a tendency to be among the most highly expressed genes according to the bulk RNA-seq data. 'Ambiguous' genes span a wider range of expression values that were between the values observed for the 'conserved' and genes that were called absent in all species (low; figure 22). 'Not analysed' genes group lies in RPKM order of magnitudes between 'ambiguous' and 'low' category genes. The more cell types affected by the expression difference of the same class, the clearer the signal in the bulk data. Additionally, I sought to investigate, whether the genes in different classes exhibit any biases in their chromosomal locations. For this, I counted for each chromosome in mouse the number of genes in different classes (figure 22.D). Ignoring the contigs that were not associated to a chromosome (very low number of genes), all chromosomes exhibit comparable distributions of genes from different classes (figure 22.D).

Since each gene could be assigned to a different class in different cell types, I quantified specific and shared classifications for the above mentioned subset of cell types (figure 23.A & figure S8). The highest number of cell type exclusive 'human gained' genes could be attributed to oligodendrocytes (47 genes), followed by Purkinje cells (45), UBCs (38) and astroglia (29). The highest number of shared 'human gained' genes was five, shared between oligodendrocytes and astroglia. The 'human lost' class had its most abundant cell type specific classifications in UBCs (22), followed by glutamatergic deep nuclei cells (19), GABAergic deep nuclei cells (18), and interneurons (11). Six genes, shared between Purkinje cells and GABAergic deep nuclei cells, were the most abundant intersect between cell types in this class. The number of cell type exclusive 'Mouse gained' genes is the highest for granule cell (20), oligodendrocytes (19), and astroglia (15). UBCs aggregated the highest number of exclusive "mouse lost" genes (65), followed by oligodendrocytes (44), astroglia (22), and granule cells (45). The number of 'Marsupial expressed' genes is the highest in UBCs (39), oligodendrocytes (30), and interneurons (22). 73 genes were classified as "eutherian expressed"

in oligodendrocytes, 37 in astroglia and 35 in granule cells. To conclude, the majority of the presence/absence expression differences are called in a single cell type.

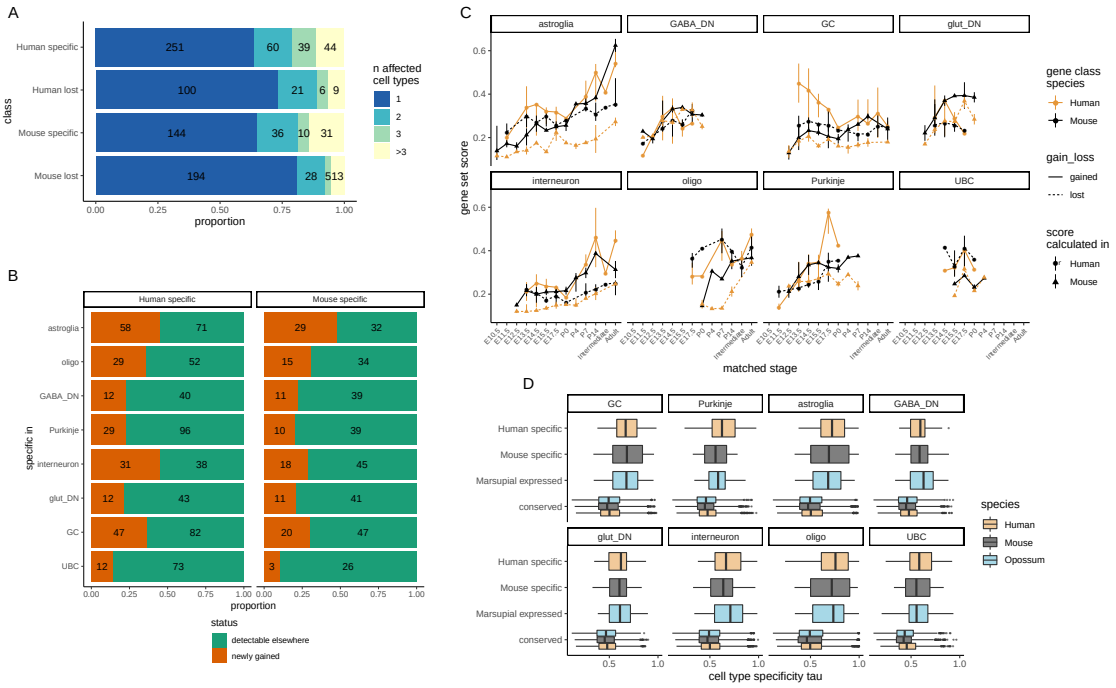


Figure 23: Gained and lost genes characteristics A: number of affected cell type within each gain/loss class. Absolute numbers are printed, proportion per class is plotted. **B**: number of genes per human and mouse-specific class per cell type of genes which are either exclusively in the observed cell type expressed, or were detected in any other cell type. Absolute numbers are printed, proption per class and cell type is plotted. **C**: gene scores per class determined per cell type for human and mouse. Scores were calculated per biological replicate and median (point) as well ass minimum and maximum observed expression (linerange) was calculated. For lost genes (dashed line), the expression was estimated by values taken from the other species (lost in human = expression in mouse, lost in mouse = expression in human). Biological replicates were assigned to matched timepoints. **D**: cell type specificity [98] per gene per class per cell type was determined for either species-specific or conserved genes.

Next, I asked whether the high number of cell type specific classifications might indicate generally higher cell type specificity of the genes affected by presence/absence expression differences. I calculated the specificity metric called tau [98] at the level of cell types in the snRNA-seq datasets. The genes with species-specific expression in different cell types have higher tau than the conserved genes, confirming my hypothesis. (figure 23.D)

I then investigated the temporal patterns of the genes with presence/absence expression differences. For each class I calculated a gene score (as done earlier) that summarizes the expression of all genes in the class, and studied the summarized expression in different cerebellar cell types across developmental stages (figure 23.C) For the gained (species-specific) genes, the score was calculated using the expression values within the respective species. The score for human lost genes was calculated using the mouse data, and *vice versa* for the mouse lost genes. For the majority of cell types, the expression scores of genes with gained or lost expression in human or mouse increase during development. Interestingly, this pattern does not hold true for genes that gained expression

in human granule cells. Specifically, the expression of human-gained genes is high in early granule cell development, declines till matched timepoint P0, and stabilizes during postnatal development. Taken together, with a few exceptions, the genes with presence/absence expression differences tend to have higher expression in late developmental stages.

Next, I sought to identify 'human gained' and 'mouse gained' genes that are not expressed in any other cerebellar cell type. Using the criteria to call presence and absence of expression, I quantified how many of the genes that gained expression in a cell type were called as present in other studied cell types (figure 23.B). The results of this analysis show that the majority of genes with expression gains are expressed in other cerebellar cell type(s) within the same species, and only a minority of thes genes can be considered as newly recruited to cerebellar transcriptomes.

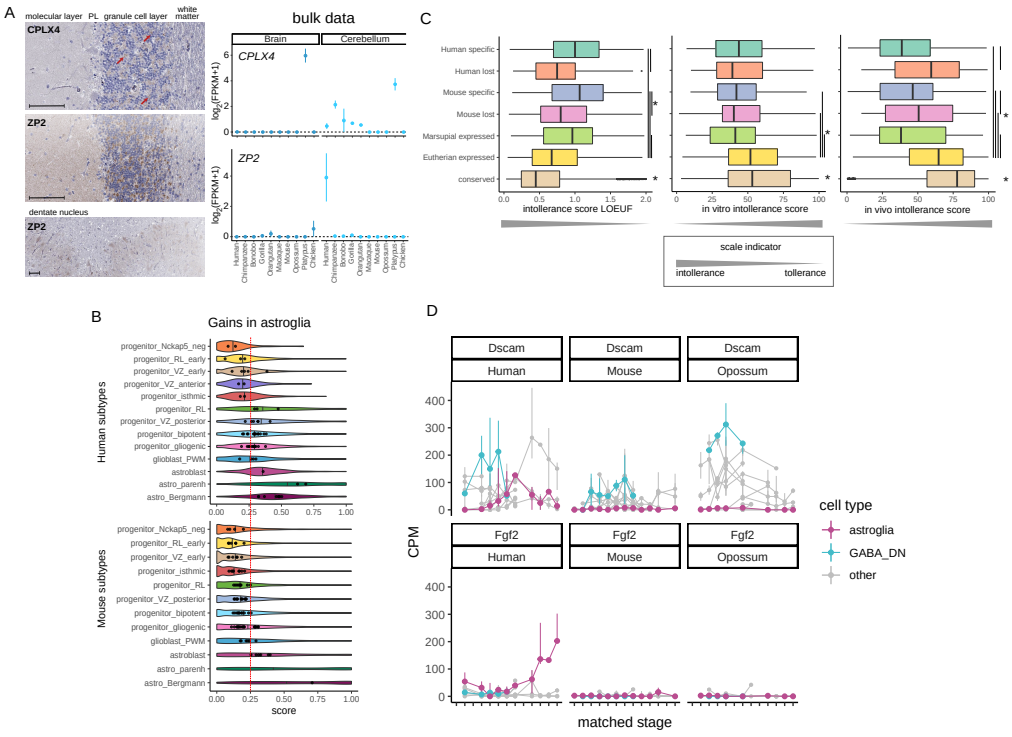


Figure 24: Evolutionary characteristics of gains and losses A: Left: *CPLX4* and *ZP2* protein signal in either the cerebellar cortex (upper two), or dentate nucleus (lower, *ZP2* only) (Human Protein Atlas [99]). Right: bulk RNA-seq determined expression of genes coding for either protein in human, chimpanzee, bonobo, gorilla, orangutan, macaque, mouse, opossum, platypus, and chicken within the cerebellum or cerebrum (brain). Mean expression (point) and minimum/maximum are shown (errorbars). **B**: gene score for expression of genes classified as either human (upper), or mouse (lower) gains in astroglia cell type grouped by subtype. Dots represent individual biological replicate pseudobulks per subtype. The red line shows the assigned cut-off of 0.25, separating "high" and "low" gene scores. **C**: metrics of functional constraint per gain/loss class across all investigated cell types. **D**: examples of human gained expression in astroglia cells (*Fgf2* and *Dscam*) shown in all three species across all studied cell types. Astroglia and GABAergic deep nuclei ($GABA_{DN}$) are color-coded, the other cell types are shown in gray. Pseudobulks were generated to estimate the expression per individuum, cell type and matched timepoint. Points represent median expression, errorbars minimum and maximum detected expression.

Two genes, which were newly recruited to a cerebellar cell type trancriptome in human, and

not detected in the other species, were *CPLX4* in interneurons and *ZP2* in granule cells. Using the immunohistochemistry data from the Human Protein Atlas, we found that the spatial distribution of protein expression signals fit with the cell type predicted from the snRNA-seq data (figure 24.A). Attempting to time the occurrence of these genes expression in the primate lineage, we used published bulk RNA-seq data Brawand *et al.* [100], to map their expression in the adult cerebellum of six primates, mouse, opossum, platypus, and chicken. *CPLX4* expression was clearly detectable in the cerebella of human, chimpanzee, bonobo, gorilla, and platypus. For comparison, in the cerebrum *CPLX4* expression was only detected in platypus. In contrast to that, the only species which showed strong expression of *ZP2* in the cerebellum was human. These results suggest that cerebellar expression of *CPLX4* emerged in the great apes lineage, and *ZP2* in the human lineage.

Since many human and mouse gains were attributed to astroglia cell type (including progenitors and astrocytes), I zoomed in on the identified subtypes, and calculated the gene expression score for the gained genes in these groupings for human and mouse. Comparing both species, it became apparent that human progenitor subtypes of posterior VZ and RL exhibited higher scores, compared to their mouse counterparts. When conducting differential gene expression analysis between the astroglia subtypes in human, 19 human-specific genes enrich in both aforementioned subtypes (table 8, hypergeometric test, $p < 0.01$). These genes represent interesting targets for downstream spacial transcriptomic and functional work to understand human specific astroglia behavior. Furthermore, these genes might be directly connected to the recently identified human-specific ventricular zone secondary progenitor pool [5].

Table 8: human-specific expression in VZ subtypes

Gene symbol [Human]
SLC16A
CORO2A
GRM1
DUOX1
SCN11A
PIEZO2
PLCZ1
GHRL
ACAP1
LEAP2
EPB42
KCTD19
LMOD3
ZC2HC1C
HES2
CPS1
BOLL

Further characterization of the genes with presence/absence expression differences was done by observing these groups in context of previously introduced intolerance scores (figure 24.C).

Intolerance to functional mutations in human populations (LOEUF) is more relaxed for any classified specific gain or loss group, compared to genes, expressed in a given cell type in all therian species (conserved). Similar signals were also present for *in vivo* and *in vitro* scores. It is to note, that for *in vivo* scores, gained expression classes were always significantly relaxed in tolerance than their lost class counterparts.

In sum, these analyses reveal numerous candidate genes that may contribute to the phenotypic differences in the cerebella from different mammalian species. The here selected genes could be used in further functional assays to investigate their contributions within the specific cell type and species.

4 Discussion

To gain insights into gene expression dynamics during development and evolution, studies analysing large bulk RNA-seq datasets have been conducted (e.g. [7, 66, 100]) Even though the high sequencing depth in bulk RNA-seq datasets enables capturing the variability of expression profiles across development and evolution, the detected signals represent average gene expression profiles across all cell types in the studied tissues, and are therefore influenced by the changes in the relative abundance of the different cell types in the studied tissues. Thus, the observed differences in gene expression levels, for example between species, cannot be unambiguously attributed to true expression changes in the specific cell types in the tissues. Methods like FACS sorting can help to resolve this issue by allowing enrichment of specific cell types based on marker genes [101], however these approaches are very time-consuming and require previous knowledge about the marker genes. In the last years, single-cell transcriptomics (and more recently multiomics) approaches have emerged [102]. These technologies enable studying gene expression profiles of heterogeneous cell populations in complex tissues.

4.1 snRNA-seq atlases of cerebellum development

I, together with Dr. Mari Sepp and a team at the Kaessmann lab[2], applied single-nucleus RNA-seq techniques to study the evolution and development of the mammalian cerebellum. Data for three species was generated: human, mouse (*Mus musculus*), and opossum (*Monodelphis domestica*). The generated datasets allowed me to describe the development of the cerebellum in all three species on a single cell level and with high temporal resolution starting at early neurogenesis and extending into adulthood. Similar experiments have previously been carried out to characterize cerebellum development in the mouse [78, 88, 103] and human [95], but the datasets presented here have the advantage that they allow to study cerebellum development from the evolutionary perspective. Using opossum as the outgroup for eutherian mammals, it is possible to contrast changes detected in either human or mouse, and to estimate on which evolutionary lineage the change in gene expression or in cell type abundance emerged.

To be able to draw meaningful biological conclusions from this work, I had to overcome several technical limitations associated with a change in Chromium library preparation chemistry and different quality of genome annotations between the studies species. During the data generation phase of this study, 10x Genomics updated the 3' gene expression kit chemistry from version 2 to version 3. In our data the version-dependent shift in the number of detectable genes is clearly visible: nearly double the number of UMIs per cell are detected in version 3 libraries compared to version 2 libraries. This increase in UMI numbers correlates with a higher number of genes detected. To

[2] As described in the beginning of this thesis.

estimate gene expression levels, I determined both exonic counts as well as pre-mRNA counts. The latter approach use the whole gene model as basis, and is known to increase the depth of single-nucleus and single-cell RNA-seq data [96]. Evgeny Leushkin helped to re-annotate the opossum genome to improve gene detectability. The obtained alignment metrics in the three species were very similar. Overall, the samples, when combined to pseudobulks, showed clear developmental signals and clustered accordingly. Though, differences in libraries prepared with different Chromium versions were detectable and lead to higher than expected variability in the dataset. Nonetheless, the main signal was of biological nature, outweighing the technical influences.

4.2 Annotation and species alignment

The first steps in the analyses of the snRNA-seq datasets involved integration of data across developmental stages and species, annotation of the mouse dataset, transfer of the annotations from mouse to human and opossum, and finding the cross-species correspondences between the sampled developmental stages. I used LIGER [77], an established method for batch-correction, to integrate the snRNA-seq datasets across development. Integration of complex datasets always bears a risk to remove too much signals that differentiates biological units, for example stages or cell types. I found out, that the initial LIGER non-negative matrix factorization integrates the data but does not remove all stage specific effects, if the final recommended step of aligning the datasets after factorization is omitted. I quantified the within and between stage Euclidean distance, and even though the cells set closer together in the LIGER embedding, stage specific signals are still present. This led me to conclude that major developmental signals are preserved in the integrated dataset. LIGER was also used to integrate data across species, however further correction using MNN correct [75] was required to achieve well integrated datasets in low dimensional space.

After iterative clustering of the integrated mouse dataset and identification of the marker genes, Dr. Mari Sepp and me annotated the cell types in the dataset as detailed as possible, using literature and online resources like the Allen Developing Mouse Brain Atlas and GenePaint [79, 80, 99]. A Hierarchical annotation model was applied: cells were assigned to a broad developmental lineage, cell type, cell differentiation state and, if possible, to a subtype. The annotations in the mouse dataset were then transferred to the human and opossum datasets using the pairwise integrated embeddings and manually curated to allow identification of cell types or states which were not found in the mouse dataset. Using this procedure we were able to detect all cell types expected to be present in the developing cerebellum [4, 88, 95]. In terms of resolution the annotations in the current study exceeded the previous studies on mouse [88] and human [95], as also described in the following sections.

When development of an organ is studied in evolutionary context, developmental correspondences need to be established prior to cross-species comparisons [7, 66]. The species studied here

exhibit vast differences in developmental tempo. To establish correspondences between the samples developmental stages, I measured correlations beween transcriptomal profiles, pseudoages [82], and cell state proportions. All three measures show a high degree of agreement, which also overlaps with the bulk RNA-seq data [7]. Thus, no major heterochronies between the species can be reported.

4.3 Cellular composition and abundance dynamics

Cerebellar development is a very dynamic process [4, 67, 95, 104]. The data of this study reflects this. At level of the main neural lineages in the cerebellum, we identified in all species: glial cells, GABAergic neurons originating at the ventricular zone, and glutamatergic neurons originating at the early or late rhombic lip.

The RL associated lineages comprise glutamatergic cells: isthmic nuclei cells, glutamatergic deep nuclei neurons, granule cells and unipolar brush cells. Granule cells are the most abundant cell type in the brain and make up more than 80% of the adult cerebellum [2]. After initial specification, granule cell progenitors (GCPs) accumulate in the external granule cell layer where secondary proliferation occurs [51]. Cells at the GCP state were detected in all three species, clearly showing granule cell progenitor characteristics and cell cycle activity. A related subgroup of cells, denoted as GCP/UBCP, showed expression profiles that shared features with both granule cells and unipolar brush cells. It remains to be determined in future studies if GCP/UBCP cells represent true multipotent progenitors that give rise to the two neuron types, or a mixture of transcriptionally similar unipotent progenitors. Among differentiating granule cells early and late cells could be distinguished. An additional group of differentiating GCs in mouse and opossum was distinguished on the basis of *KCNIP4* and *OTX2* expression. Although this group was not distinguished in human, it is likely explained by sampling differences. Granule cell diversity in the adult cerebellum was recently reported [9], and it was proposed that different subtypes of GCs arise at distinct developmental timepoints. Comparing the profiles of the differentiating GC subtypes with the profiles of adult GC subtypes, this notion can be confirmed. Therefore, the analysis of the data in this thesis helps to explain the sources for variety of granule cell subtypes in the adult cerebellum. Unipolar brush cells could be separated into two subgroups: one expressing the canonical marker *EOMES* [9] the other, previously not reported, shows only low *EOMES* expression, but is $HTCRTR^{+}$. This novel subtype of UBCs was detected in all species and according to the in situ hybridization data [80] locates in the internal granule cell layer in early postnatal mice. Further studies are needed to elucidate the functions of this UBC subtype in the cerebellum, but initial investigations by Dr. Lena Kutscher have already confirmed the UBC-like morphology of these cells.

Astroglia cell type lineage includes neural progenitors and astrocytes. Progenitors show clear cell cycle activity and form a continuum of subtypes representing different germinal zones in the

cerebellum. I tried to dissect the progenitors in more detail by removing the cell cycle associated genes from the analysis, but this approach was not successful. This might be due to the fact, that the cell cycle influences the whole transcriptome and even when the directly associated genes are removed, the signal of the ongoing cell division is still imprinted into the observable transcriptome. In addition to this observation, a group of cells sharing characteristics between astroglia and oligodendrocytes was detected in opossum and human. These cells might represent preOPCs, fitting with their expression of *EGFR* [82]. In mouse this cell group was not detected, which could be explained by the non-cerebellar origin of most oligodendrocytes in mouse [105].

The emergence of GABAergic cell populations in the all three datasets follows the expected temporal pattern [4] with GABAergic deep nuclei neurons being produced early in development, followed by Purkinje cells and interneurons. In all species, five interneuron subtypes were distinguished, marked by distinct marker genes (*ZFHX4*, *RGS6*, *KLHL1*, *NXPH1*, *SORCS3*). Four of these subtypes were directly matched to recently described adult interneuron subtypes in the different layers of the cerebellar cortex data [9]. In opossum, an additional group of interneurons could be identified (*MEIS2*). It is currently unknown, if these cells are contaminating cells from other brain regions, or a more rare interneuron subtype in opossum. We identified four Purkinje subtypes in mouse and opossum, which can be differentiated by *EBF1* and *EBF2* combinatorial expression profiles, with *RORB* (*EBF1* high, *EBF2* low), *CDH9* (*EBF1* and *EBF2* high), *ETV1* (*EBF1* low and *EBF2* high) and *FOXP1* (*EBF1* low, *EBF2* high) as markers. These subtypes show distinct spatiotemporal patterns in the developing mouse cerebellum. It is known that adult Purkinje cells show spatial patterning in adult cerebellum and form parasagittaly arranged stripes of ALDOC-positive and negative cells [21]. Comparisons to recently described adult Purkinje subtype transcriptomes [9] revealed correspondences between the developing and mature subtypes, thus adding a spatial component to the previously postulated temporal origin [41, 42, 46, 106–108] of the patterning of Purkinje cells. Chung *et al.* described EBF2 as repressor of ADLOC positive phenotype previously, which fits to the observations made in this study. In line with these findings, genes with variable expression across Purkinje subtypes are enriched for cadherin family genes, confirming the importance of cadherins in Purkinje cell patterning [109]. Notably, human developing Purkinje cells did not show the same subtype composition as seen in the mouse and opossum. The two identified human subtypes differentiated by birth date, and were either *EBF1* and *EBF2* low (early born) or high (late born). Whether this difference between the species is due to technical variations in sampling, and/or of biological origin remains to be investigated. Another notable observation for Purkinje cells was their approximately two fold higher relative abundances in human compared to mouse and opossum at the developmental stages when Purkinje cells are generated in the cerebellum. I created a Bayesian hierarchical model to test whether the observed difference could be explained by random events. I chose this approach to be able to leverage the biological replicates present in

the data, but allowing variability in measured abundances per biological sample. Furthermore, the Bayesian approach allowed me to estimate the mean per species and stage as a hyperparameter, which could not directly be measured but needs to be inferred. And finally, Bayesian models don't suffer from multiple testing, as well as using the region of practical equivalence (ROPE, reviewed in [110]). By determining the highest density interval for the modelled posterior distance, practical equivalence between species could be confirmed if it crosses zero difference, which is not possible in frequentistic statistics. Assuming the model approximates reality well enough, the difference in abundance in human could be attributed to a true increase in developing Purkinje cells. Whether the difference could be connected to recently described basal progenitors [5] in human is yet to be elucidated.

These results show that the overall cellular diversity in the developing cerebellum has been conserved for at least 160 million years. Shifts in relative cell type abundance are shared between the three studied species. The most obvious species-specific deviation can be observed in developing human Purkinje cells, which double in proportion during two consecutive stages. Comparing the developmental data with published adult mouse cerebellar cortex data allowed to match adult patterning to developmental programs, thus adding valuable information about the process that gives rise to specialized areas of the adult cerebellum.

4.4 Conservation of gene expression programs

The overall cellular dynamics during cerebellar development is conserved between the species, as discussed above. Next, the degree of gene expression conservation was investigated. I propose that genes that have conserved expression profiles between the species are key players in defining cell type identities. A broad similarity of gene expression patterns was studied by principal component analysis of cell state pseudobulks. This analysis demonstrated that the overall transcriptomic landscape of cell type programms is conserved throughout the studied species. Even though the naive PCA showed species-specific components appearing early on, the first principal component, capturing about 80% of all variance in the dataset does not show species separation. This component captured the signal of development which must therefore be largely shared. The centered expression matrices remove the species specific signals within the first ten principal components and captured the variety of cell types, species agnostic, similar to the bulk RNA-seq based analysis of organ development [7].

I identified the conserved marker genes for each cell state by calling marker genes and filtering for the ones that are shared between the three species. The majority of cell state markers called are not shared between species (i.e. not conserved), in line with previously made observations in adult mouse cortex data [97]. Nevertheless, if a marker is called in a single species, it does not always mean that the gene is not expressed in the same cell state in the other two species. Instead, in these

cases the differences often lie in the expression specificity of the gene (figure S7). However, a set of conserved markers that show the same expression specificity in all species, could be defined for all cell states in the datasets.

Interestingly, among the conserved marker genes, transcription factors are enriched. This is indicative that transcription factors play a vital role in cell type and state definition, whereas the target genes might vary between species, supporting the proposed hypothesis [91]. Comparison of the transcription factors with known marker genes also contributed to this notion, since many were reported before [4, 88–90] . Besides these known connections between transcription factors and cell types or states, novel candidate cell type-defining transcription factors were found, For instance, in GABAergic interneurons these include PRDM8 and BHLHE22, gene products of which are known to form a repressor complex functioning in pallial circuit formation [111]. SATB2, known to play a role in neocortical upper layer neuron differentiation [112], was identified as a conserved marker of differentiating granule cells. The identified transcription factor associations were supported by the SCENIC analysis that models transcription factor activity based on the single-cell RNA-sequencing data [87]. Overall, a high degree of agreement between modelled transcription factor activity the two analyses are not independent due to the fact that the SCENIC model was trained on the expression data. The main drawback of the SCENIC analysis was the lower coverage, likely due to unknown binding motifs, the limitations of the method in detecting repressor activity [87], or sparse detectability of transcription factor expression in the single-nucleus RNA-seq data. Furthermore, the SCENIC analysis performed here was only based on promoter sequences, a more detailed view could be achieved by additionally including the enhancer sequences [113], mapped, for instance, by single-nucleus ATAC-seq [67]. Altogether, these analyses generated a shortlist of transcription factors with potentially important roles in the specification of cerebellar cell types during development. The roles of these transcription factors could be studied in more detail in functional experiments.

4.5 Differentiation programs in granule and Purkinje cells

Single cell techniques not only open up the possibility to identify new cell types or states, but also allow modelling of complex differentiation processes. In this study, I modelled the non-bifurcating differentiation pathways of Purkinje cells and granule cells, two important neuronal cell types in the cerebellum. The high-resolution sampling strategy of this project facilitated the capturing of cells committed to either cell type lineage throughout the differentiation process up to (GC) or close to (Purkinje cells) maturity. Using cell type specific data integration and diffusion pseudotime [114], I fitted a differentiation vector through the low dimensional embedding to model continuous changes in the gene expression space. Comparisons of the pseudotime distributions across developmental stages (real time) revealed clear differences in the modes of differentiation between Purkinje cells and granule cells, in agreement with previous knowledge. Specifically, for granule cells changes is

pseudotime values across developmental stages are gradual. This is in line with the protracted mode of GC neurogenesis that covers a substantial period of pre-and postnatal development [51, 55, 115, 116]. In contrast, Purkinje cells show sharp changes in pseudotime values between stages, reflecting a pulse of Purkinje cell production and subsequent differentiation [117, 118]. These observations suggest that the pseudotime models capture the neuronal differentiation processes accurately.

Identification of genes that show dynamic expression during differentiation of granule cells or Purkinje cells revealed a substantial overlap between the dynamic genes between the neuron types. This points to shared neuronal differentiation mechanisms. The shared dynamic genes are under stronger functional constraints than genes dynamic in one cell type only, or genes that are not dynamic. This observation is in line with studies linking genes with higher pleiotropy to more severe phenotypes [7, 66]. Moreover, genes associated with cerebellar malformations, intellectual disability, spinocerebellar ataxia or medulloblastoma are enriched in the group of shared dynamic genes, further confirming the functional relevance of these genes. Genes which are dynamic in Purkinje differentiation only are additionally enriched in genes of autism spectrum and intellectual disorder. Together, these results indicate that many of the diseases, associated with cerebellum development, most probably affect not only a single cell type but rather influence the development of the organ as a whole.

To describe the most common gene expression trajectories during Purkinje cell and granule cell differentiation, and to compare orthologous gene trajectories across species, I used fuzzy clustering. Among the genes that show conserved profiles across species (i.e., orthologous genes were assigned to the same cluster), transcription factors are enriched. For both cell types, I identified core transcription factors active in early, mid or late differentiation. For some of these transcription factors, spatial expression patterns could be traced in public ISH data [80]. This allowed replication of the known migration patterns of granule and Purkinje cells during the differentiation process, thus verifying the chosen approach.

Comparing the expression trajectories between human, mouse and opossum, and using the latter as an outgroup to eutherian mammals, I identified genes that show human- or mouse-specific expression trajectories in granule cells or Purkinje cells. Interestingly, the number of genes with human-specific trajectories is significantly higher than expected (136 human-specific vs 54 mouse-specific), whereas similar numbers of genes with trajectory changes were identified for Purkinje cells (47 human-specific, 42 mouse-specific). Only a very low number of genes (1 to 4) exhibited changes in trajectories in both cell types. This indicates, that despite the high number of dynamic genes shared between the cell types, evolutionary alterations happen rarely in multiple cell types at once. This is in line with previous observations at the level of organs Cardoso-Moreira *et al.*: "Notably, although genes with trajectory changes are broadly expressed, the changes themselves are organ-specific. Trajectory changes are restricted to one organ in 93–96% of the cases. This

is consistent with the underlying mutations affecting regulatory elements, which control a subset of the total spatiotemporal profile of each gene, and with evolutionary theory, as mutations that affect several organs are less likely to fix in populations" [7]. Similarly, the more cell types affected by a change, the higher the risk of detrimental outcomes and hence lower probability of fixation in populations. Additionally, genes with species-specific trajectories in Purkinje cells or granule cells are under weaker functional constraints than genes with preserved trajectories. This suggests that the preserved genes likely represent genes with important functions in neuronal differentiation. Expression alterations of these genes will most likely be detrimental, thus less likely to fix in populations.

Looking at single genes, SNCAIP is the most diverged human-specific gene in granule cells. Compellingly this gene is known to be frequently duplicated in medulloblastoma group 4 [119]. This specific medulloblastoma subtype has remained hard to model in the mouse [120]. The difference between human and mouse SNCAIP expression trajectories in granule cells could be one of the reasons that complicates imodelling this tumour subgroup in mice. Additionally, two genes associated to autism spectrum disorder, MYTL1L and KANSL1, show human-specific trajectories in granule cells, as well [121, 122]. These examples highlight the importance of comparative studies in informing research on the disease mechanisms in model species.

In sum, the detailed caharcterisation of the transcriptomic landscape of Purkinje cell and granule cell differentiation provides a shortlist of candidates important in neuronal differentiation for functional work, identifies the core gene expression programs in neuronal differentiation on the basis of evolutionary conservation, and informs on genes with lineage-specific expression trajectories, which may underlie phenotypic differences between the species.

4.6 Fundamental expression differences

I further assessed if fundamental/radical differences in cell type-specific transcriptomes are present between the studied species by invastigating gains and losses of gene expression. Classification of orthologous genes into the groups of conserved, gained or lost in each cell type was a complicated endeavor. Variations in clustering, cell type identification, differences in Chromium version and numbers of cell were not easy to circumvent. The chosen approach was designed to be conservative rather than exhaustive. Stringent cutoffs of expression levels and fold changes were used, as well as prior information from bulk RNA-seq data [7], and internal expression controls. Even though, all of these measures were applied, technical artifacts cannot be entirely ruled out. Therefore, I would like to stress that this analysis aims to enrich for genes that gained and lost expression in the different cell types in a species , but does not prove either of it for individual genes without further experimental validation. This is also reflected in the number of genes classified as ambiguous or not analysed, which were by far the biggest groups.

The highest numbers of genes, besides the conserved and aforementioned classes, are in the classes Eutherian expressed and human gained. The smallest group is "human lost". I tried to verify the classification by using bulk RNA-seq data to test, whether genes that are lost or gained species-specifically at the cell type level, also exhibit expression differences at the bulk level. A general trend in the expected direction was observed: the more cell types are affected by a gain or loss, the clearer the signal in the bulk data. Though, exceptions are also detectable, most probably driven by expression deviations within the granule cell lineage. Since this is the dominating cell type in the postnatal cerebellum, changes within this group of cells are more easy to pick up in the bulk RNA-seq data. Gained or lost genes are mainly classified as such only in one cell type, though this does not mean that their expression is specific to that cell type. This is very similar to the observation made for trajectory changes, where a given gene is dynamic in both studied cell types but only altered in one. Cell type specificity is higher for genes with species-specific expression gains, indicating that expression changes are more likely to occur in genes with less pleiotropic expression profiles. However, I found that it is more likely to gain expression of a gene in a given neural cell type if it was expressed in another neural cell type before, possibly due to similarities in gene regulation between neural cells in general. By studying aggregated expression profiles of genes with expression gains or losses, I found that expression levels are in general higher at later stages of development with one exception: human gains in granule cells. The latter genes show high expression in at early stages. This might indicate a specific alteration of GCP transcriptomic landscape in the lineage leading to humans.

Regarding functional constraints, genes with conserved expression (i.e. called present in all species) show highest constraint, and genes that exhibit mouse or human-specific expression profiles the least. Among the latter, genes with expression losses in specific species are under higher constraints than genes with expression gains. This could be a hint that gains of expression are evolutionary speaking cheaper to explore than losses, due to the high number of interconnections of a present gene and the less tight and yet to be established regulation of a newly expressed gene.

Among the genes, which were identified to be human gains, were *CPLX4* (interneurons) and *ZP2* (granule cells). According to immunohistochemistry data [99], both gene products are distributed with expected patterns in the human cerebellum: *CPLX4* is detected in granule cell interneurons and *ZP2* in granule cells themselves. *CPLX4* has been shown to be expressed in the mammalian retina and functions in synaptic vesicle exocytosis [123]. *ZP2* (zona pelucida 2) was previously shown to be expressed in the human cerebellum but not in the cerebella of two other primates [124]. Using the bulk RNA-seq data from Brawand *et al.* [100], we found that *ZP2* expression in the cerebellum is unique to humans, whereas *CPLX4* is also expressed in the cerebella of other great apes. This analysis helps to hone in when during evolution the gain in expression happened. Additionally, observing human gains in astroglia lineage, high expression levels were observed in

posterior ventricular zone and rhombic lip progenitors, when compared to the expression level of mouse gains in the same subtypes in mouse. Among the genes with gained expression in astroglia, 19 were detected as enriched in posterior VZ and RL progenitors. It could be speculated that these gains of expression might be connected with the higher relative abundances of developing Purkinje cells in human, as well as the recently identified basal progenitors in the human cerebellum [5].

Among the gene with expression gains in human are several known disease-associated genes. For instance *FGF2*, linked to pilocytic astrocytoma, and *DSCAM*, which is connected to Down syndrome, gained expression in human astroglia. *FGF2* shows clear human astroglia specificity and is not expressed in other cell types in the cerebellum. In contrast, *DSCAM* is expressed in other cell types, shared between the species, for example in GABAergic deep nuclei neurons. Human-specific expression patterns of disease associated genes can highly impact clinical research and need to be investigated further.

Taken together, the analysis of presence and absence of expression allowed enrichment for genes, which are potentially gained or lost in a given cell type in a species-specific manner. The results are in line with previous analyses in this thesis, highlighting a generally conserved framework of cerebellar development but also anraveling lineage-specific alterations.

5 Conclusion

Using the state-of-the-art single-nucleus RNA-seq technologies, combined with a high resolution developmental and evolutionary dataset allowed the study of the mammalian cerebellum in unprecedented detail. The development of the mammalian cerebellum is a tightly regulated process, which leads to immense cell proportion shifts during its maturation. These patterns and fundamental gene expression programs are conserved between eutherian and marsupial species separated by 160 million years of evolution. Key cell type and state-specific markers could be identified, which can be the basis for further research that focuses on the functional characterization of specific groups of cells. Even though cerebellar development is highly conserved, changes in gene expression in individual cell types were detected. Some of the genes with altered expression profiles are associated with neurodevelopmental diseases or cancer, indicating that for these genes not all aspects of the disease can be modelled in the mouse, assuming the associated genes truely contribute to the disease phenotype. Key concepts of evolutionary research could be confirmed on a cell type level: changes in gene expression often affect single cell types, even if a given gene is expressed in a variety of cell types and states.

6 Material & Methods

6.1 Sample preparation and data generation

My PhD project focused on the analysis of the presented data. I did not contribute to any wet-lab work which was the basis for any data that was present in this thesis. Please refer to our preprint [10] for any details about experimental procedures like dissections, tissue dissociation, library preparation and sequencing settings. All of these experiments were conducted by Dr. Mari Sepp with support of Noe Mbengue, Celine Schneider and Julia Schmidt. Data was produced using 10x Chromium 3' chemistry versions 2 and 3.

6.2 Alignment reference generation

The basis of all genome annotations used in this work was ENSEMBL version 91. Mouse (mm10) and human (hg38) assemblies and associated genome annotations were retrieved. For opossum ENSEMBL version 87 annotations were extended with stranded poly-A bulk RNAseq based predictions[7] using a previously reported approach[125]. The predictions of additional features were done by Evgeny Leushkin. Cellranger (10x Genomics) references were generated with default settings using the `cellranger mkref` (v3.0.2) function with default settings. Opossum chromosomes 1 and 2 were too big to be processed by cellranger, hence I split both chromosomes at position 536,141,000 after verifying that no annotated feature is disrupted.

6.3 UMI countmatrix generation and data preprocessing

Sequencing results were demultiplexed using `cellranger mkfastq` (v3.0.2) and aligned with `cellranger count` to the appropriate genome. This generated not only alignments but also unfiltered UMI count matrix for all detectable barcodes. Annotations for either full gene models (pre-mRNA), or exon only were provided and each dataset quantified with each set of features. If not stated otherwise, we used the pre-mRNA counting for the majority of analyses.

Loaded barcodes were identified by calculating the fraction of intronic UMIs of all detected UMIs. Once, the fraction was calculated by dividing the difference of pre-mRNA and exonic counts by the pre-mRNA counts, a Gaussian mixed model (k=2) was fitted (`mclust`, v5.4.3, [126]) to these values. The cluster with the highest average intronic UMI fraction was chosen as valid barcodes. Potential doublets were removed using `Scrublet` (v0.2, [127], python 3.6) by calculating the doublet scores for all valid barcodes and removing the barcodes above the 90%-percentile.

In total 115,282 mouse, 180,956 human and 99,498 opossum nuclei passed all previous filters and were subjected to the downstream analyses[3].

[3] Except a single human sample, which was later identified to contain a high fraction of contaminating neighboring tissue, see subsection "Data Annotation" for details.

6.4 Quality control and sanity checks

To infer general data quality, reproducibility and sanity, data of all nuclei of any given sample were aggregated, only genes which showed expression in at least 10% of cells in at least one batch were considered. Spearman correlation calculated (data not shown). Samples of the same developmental stage showed high correlation (> 0.75), even between different chromium versions.

6.5 Per species data integration and clustering

A major obstacle in the here presented analysis was to overcome batch effects, Chromium version influences and developmental signal induced transcriptomic shifts, to identify cell types and states, without removing differentiation signals. Various methods and approaches were investigated and in the end I settled with LIGER[77] (v0.4.2). Batch annotations were provided without further definition of stage or biological replicate attributes. LIGERs standard approach for normalization and highly variable gene detection was applied, followed by integration using `optimizeALS` function with parameter $k = 100$. The resulting per-batch non-negative components were combined and subjected to uniform manifold approximation projection, UMAP [71], using the R uwot (parameters: `n_neighbors = 15`, `min_dist = 0.15`, `metric = "cosine"`). [128] package (v0.1.10).

To resolve the complexity of the datasets at highest possible resolution, I chose an iterative clustering approach: first, I applied the Louvain community detection algorithm [129] as implemented in SCANPY [74] (v.1.5.1) with the resolution parameter set to 3. Each cluster then underwent the same treatment as the full dataset: LIGER integration ($k = 25$) and a second round of Louvain clustering. Batches which contributed less than 50 cells to the cluster were excluded from the LIGER integration. The number of top-level clusters were: 68 in human, 61 in mouse and 67 in opossum. The number of low-level clusters was approximately 600 per species. Marker genes were identified for all clusters using the `quickMarker` function of the SoupX package [130]. This function uses TF-IDF transformation (term frequency - inversed document frequency) for specificity scoring and hypergeometric test for enrichment p value estimation. Markers were selected by ranking the genes by TF-IDF values and a p-value cutoff of $\alpha = 0.01$.

6.6 Cross-species integration

To leverage the high amount of mouse-cerebellum-specific literature for annotating the other two species, I decided to integrate all three species in one common high dimensional embedding. This allowed me to transfer the mouse annotation to human and opossum (see following sections). After testing various strategies, the following proved itself to be the most effective: first, all species were integrated using LIGER, to which I provided only the batch information per cell. This step projected the data into one common embedding with 100 dimensions. UMAP investigations of this embedding

revealed that the data per species was ntegrated, overcomming developmental and batch signals, but species-specific clustering dominated the embedding. Hence, the next step was to correct the non-negative matrix factorization coordinates using MNN correct [75] (`fastMNN` function of the batchelor package, v1.0.1). This resulted in an embedding which overcame most of the batch-, differentiation, and species-specific-effects. The same strategy was used to perform pairwise integrations between mouse and human and opossum and mouse.

6.7 Data annotation

The integrated mouse dataset was the first to be annotated, due to the high amount of literature concerning mouse cerebellar structure, functions and development. Marker genes were called per cluster[4] and the resulting gene lists underwent thorough literature research by Dr. Mari Sepp and me. Additionally to literature resources, we used publicly available databases: Allen Brain and Developing Brain Atlas [79, 80] and the Human Protein Atlas [99]. The annotation strategy Dr. Mari Sepp and I developed based on a hierarchical annotation structure is defined as follows: Clusters were assigned to a broad lineage (ventricular zone, rhombic lip+, neuroepithelium and mesoderm). Per lineage cell types were identified and cell types were split into cell states, reflecting differentiation maturity. If residual variance was detectable, cell states were further subdivided into subtypes, which might reflect spatiotemporal specifications of the observed cell states. The hierarchical annotation scheme is shown in table 10. The following exceptions were made to this scheme: "GC/UBC" cell type was implemented, due to the mixed signal in these clusters, showing markers for granule cells and unipolar brush cells. Often, early neuroblasts could not be resolved in regards to their cell fate, therefore the catch-all cell types "VZ neuroblast", or "NTZ+ neuroblast" were created, describing post-mitotic cells. Subtypes of astroglia cells often reflect spatiotemporal groups which are not necessarily different subtypes but can originate from the same cell state. Oligodendrocyte differentiation was captured at the subtype level. And finally, all immune cells were grouped into a single cell type (immune) due to low numbers.

To annotate human and mouse, I used the pairwise integration with the now annotated mouse dataset: within the common embedding each cluster was aggregated to its centroid within the 100 dimensional embedding. This centroid was used for correlation coefficient calculations (Pearson) between the species. Highest correlation coefficient between any human/opossum and mouse cluster lead to initial mouse to human/opossum label transfer. Using the correlation coefficient as a measure of confidence, all transferred labels were manually validated using published literature. Table 9 summaries the results.

Furthermore, within the human dataset, after initial human-specific integration, one dataset (SN296) showed expression of *HOX* genes, which were not expected in the cerebellum, within two

[4]For either toplevel or lower level clusters.

Table 9: Dataset annotation status per species

	Human	Opossum	Mouse
Match verified	69%	51%	
Re-annotated (sampling difference)	4.2%	9.9%	
State / subtype adaption	23%	19%	
Cell type adaptation	1.7%	4.7%	
Remove (unclear identitiy)	2.8%	4.8%	
Total:			
Annotated after curration	98%	94%	97%
Subtype assigned	47%	40%	51%
Unlabeled	1.2%	3.7%	2.5%

distinct clusters (cluster 23, 34). Hence, this single batch was removed for all further downstream analyses.

Table 10: Levels of data annotation

origin	cell type	cell state	subtype	species
mesoderm	erythroid	erythroid		Human/Mouse
mesoderm	immune	immune	microglia	Human/Mouse/Opossum
mesoderm	immune	immune	nonparenh macrophage	Human/Mouse
mesoderm	immune	immune	T-cell	Human
mesoderm	immune	immune		Human/Opossum
mesoderm	mural/endoth	mural/endoth		Human/Mouse
NE	astroglia	astrocyte	astro Bergmann	Human/Mouse/Opossum
NE	astroglia	astrocyte	astro parenh	Human/Mouse/Opossum
NE	astroglia	glioblast	astroblast	Human/Mouse/Opossum
NE	astroglia	glioblast	glioblast PWM	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor bipotent	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor gliogenic	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor isthmic	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor MB	Opossum
NE	astroglia	progenitor	progenitor *Nckap5* neg	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor RL	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor RL early	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor RP	Mouse
NE	astroglia	progenitor	progenitor VZ anterior	Human/Opossum
NE	astroglia	progenitor	progenitor VZ early	Human/Mouse/Opossum
NE	astroglia	progenitor	progenitor VZ posterior	Human/Mouse/Opossum
NE	astroglia	progenitor		Mouse/Opossum
NE	ependymal	epend progenitor		Opossum
NE	ependymal	ependymal		Mouse/Opossum
NE	GABA MB	GABA MB		Human
NE	isthmic neuroblast	isthmic neuroblast		Human/Opossum
NE	MB neuroblast	MB neuroblast		Opossum
NE	MBO	MBO		Mouse/Opossum
NE	meningeal	meningeal		Human/Mouse/Opossum
NE	motorneuron	motorneuron		Mouse
NE	neural crest progenitor	neural crest progenitor		Mouse

Continued on next page

Continued from previous page

origin	cell type	cell state	subtype	species
NE	NTZ mixed			Human/Mouse/Opossum
NE	oligo	oligo progenitor	COP	Mouse/Opossum
NE	oligo	oligo progenitor	COP early	Human
NE	oligo	oligo progenitor	OPC	Mouse/Opossum
NE	oligo	oligo progenitor	OPC early	Human
NE	oligo	oligo progenitor	OPC late	Human
NE	oligo	oligo progenitor	pre OPC	Human/Opossum
NE	oligo	oligodendrocyte		Human/Mouse/Opossum
RL+	GC	GC defined		Human/Mouse/Opossum
RL+	GC	GC diff 1	GC diff 1 early	Human
RL+	GC	GC diff 1	GC diff 1 late	Human
RL+	GC	GC diff 1		Mouse/Opossum
RL+	GC	GC diff 2	GC diff 2 early	Human/Mouse/Opossum
RL+	GC	GC diff 2	GC diff 2 *Kcnip4*	Mouse/Opossum
RL+	GC	GC diff 2	GC diff 2 late	Human/Mouse/Opossum
RL+	GC	GCP		Human/Mouse/Opossum
RL+	GC			Human/Mouse
RL+	GC/UBC	GC/UBC diff		Mouse/Opossum
RL+	GC/UBC	GCP/UBCP		Human/Mouse/Opossum
RL+	glut DN	glut DN defined	glut DN posterior	Human/Mouse/Opossum
RL+	glut DN	glut DN defined	glut DN ventral	Human/Mouse/Opossum
RL+	glut DN	glut DN mature		Human
RL+	glut DN	glut DN maturing		Human
RL+	isth N	isth N defined	isth N *Nr4a2*	Human/Mouse/Opossum
RL+	isth N	isth N defined	isth N *Slc5a7*	Mouse/Opossum
RL+	isth N	isth N defined	isth N *Sst*	Human/Mouse/Opossum
RL+	isth N	isth N diff		Human/Mouse/Opossum
RL+	NTZ neuroblast	NTZ neuroblast 1		Human/Mouse/Opossum
RL+	NTZ neuroblast	NTZ neuroblast 2		Human/Mouse/Opossum
RL+	NTZ neuroblast	NTZ neuroblast 3		Human/Mouse/Opossum
RL+	UBC	UBC defined	UBC *Hcrtr2*	Human/Mouse/Opossum
RL+	UBC	UBC defined	UBC *Trpc3*	Human/Mouse/Opossum
RL+	UBC	UBC diff		Human/Mouse/Opossum
VZ	GABA DN	GABA DN defined		Human/Mouse/Opossum
VZ	interneuron	interneuron defined	interneuron early	Human/Mouse/Opossum
VZ	interneuron	interneuron defined	interneuron GL	Human/Mouse/Opossum
VZ	interneuron	interneuron defined	interneuron *Meis2*	Opossum
VZ	interneuron	interneuron defined	interneuron ML1	Human/Mouse/Opossum
VZ	interneuron	interneuron defined	interneuron ML2	Human/Mouse/Opossum
VZ	interneuron	interneuron defined	interneuron PL	Human/Mouse/Opossum
VZ	interneuron	interneuron diff		Human/Mouse/Opossum
VZ	noradrenergic	noradrenergic		Human/Mouse/Opossum
VZ	parabrachial	parabrachial		Human/Mouse/Opossum
VZ	Purkinje	Purkinje defined	Purkinje defined *Cdh9*	Mouse/Opossum
VZ	Purkinje	Purkinje defined	Purkinje defined EB	Human
VZ	Purkinje	Purkinje defined	Purkinje defined *Etv1* / *Tsx*	Mouse/Opossum
VZ	Purkinje	Purkinje defined	Purkinje defined *Foxp1*	Mouse/Opossum
VZ	Purkinje	Purkinje defined	Purkinje defined LB	Human

Continued on next page

Continued from previous page

origin	cell type	cell state	subtype	species
VZ	Purkinje	Purkinje defined	Purkinje defined *Rorb*	Mouse/Opossum
VZ	Purkinje	Purkinje defined		Human/Opossum
VZ	Purkinje	Purkinje diff		Human/Mouse/Opossum
VZ	Purkinje	Purkinje mature		Human/Mouse/Opossum
VZ	VZ neuroblast	VZ neuroblast 1		Human/Mouse/Opossum
VZ	VZ neuroblast	VZ neuroblast 2		Human/Mouse/Opossum
VZ	VZ neuroblast	VZ neuroblast 3		Human/Mouse/Opossum

6.8 Cell type abundance quantification and comparison

To quantify the cell type abundances in the developing cerebellum in human, mouse, and opossum, cells were grouped according to the assigned cell type label per developmental stage. Batch information was disregarded and assumed that cellular abundances are similar in each batch per developmental stage[5] (figure S5). Additionally, I quantified the relative cell type abundance per biological replicate. If multiple technical replicates were present, median cell type abundance was used. For this analysis, I removed the human samples focusing on deep nuclei, furthermore, I removed cell types which do not belong to the cerebellum[6] and cells which were not annotated.

To circumvent edge-cases of the clustering done on the integrated dataset[7], only cell types were considered in the aforementioned analyses, which were represented with at least 50 cells per developmental stage. This filter removed up to 280 cells per stage.

Difference in cell type abundances were modelled using a Bayesian approach by building a hierarchical model for cell type abundances (equation 1) and comparing the posterior difference distribution per stage. Biological replicates were aggregated and cell type abundances were fit to the previously mentioned model. Bayesian modelling was done using the rstan package. Highest density intervals were calculated by the `hdi` function of the HDInterval package with 95% credibility mass.

6.9 Overdispersed gene identification

Overdispersed genes were calculated as described in the results (section 3.2). In short, the data was normalized by sequencing depth via division of UMI values by the total sum of UMI per cell. Mean and variance were calculated per gene and the variance mean relationship, calculated by dividing variance by mean. The Poisson expectation was calculated by the mean of the inverse sum of UMI per cell. An arbitrary factor was added to the Poisson expectation to increase stringency. Genes

[5] High correlation coefficients between batches of the same stage confirmed, that reproducibility was given. Additionally, the cell type abundances between batches were very similar (figure S5).

[6] GABA MB, progenitor MB, progenitor isthmic, motor neuron, neural crest progenitor, isthmic neuroblast, and MB neuroblast.

[7] Such as "pulling" of cells of adjacent developmental timepoints.

which exhibited a higher variance mean relationship than the offsetted Poisson expectation were assumed to be overdispersed and therefore highly variable genes.

6.10 Pseudoage estimation

To establish stage correspondences between the species, I applied a previously published idea [82] of creating a continuous scale of developmental stages by averaging the discrete stage assignments of a given cells nearest neighborhood. Each developmental stage was assigned to its developmental rank (early to late: 1 to N). True time differences were not considered but equally sized steps between stages were assumed. Nearest neighborhood was estimated per cell by exploiting the option of the `umap` function within the uwot [128] package to return the approximated nearest neighborhood per cell. I extended this idea by not calling pseudoages species internally, but using the fully integrated embedding, containing all three species, I asked for each human and opossum cell to which stages the nearest 25 mouse cells are assigned to. Then, I calculated the pseudoage on these values, which aligned all three species, according to their highest similar mouse stage and created a continuous scale.

6.11 Establishment of stage correspondences

Due to the vastly different timings of development between human, mouse and opossum, stage alignment was vital to do comparative work on the dataset. Dr. Mari Sepp and I came up with three measures to judge stage similarity and hence identify possible stage matches:

(I) Overall transcriptomal similarity. This correlation based approach was calculated on all shared highly variable genes between either human and mouse, or opossum and mouse. Spearman's rho correlation coefficient was determined between pseudobulks resulting from the merger of all cells belonging to the same species-specific stage. Expression values were CPM normalized and mean-centered.

(II) Pseudoage similarity. As described above (subsection 6.10), shared pseudoages were determined. The resulting continuous scale was binned to get aligned stage assignments and per species-specific stage, the proportion of pseudoage bins was determined. Pairwise Manhattan distance was used as distance measure between the stages.

(III) Cell state proportion similarity. Per species-specific stage, the proportion of annotated cell states was evaluated. Pairwise Manhattan distances were calculated from the resulting proportion matrices.

To find the shortest path between human and mouse and opossum and mouse, dynamic timewarp algorithm, as implemented in the dtw package [131] (v1.20), was used. Stage correspondences were then based on the mouse developmental stages. The agreement between all three approaches was evaluated and best matching stages assigned.

6.12 Gene expression score calculation

If multiple genes expression was to be visualized and analysed, I applied an approach inspired by La Manno *et al.* [82]:

$$\begin{aligned} C_{ij} &= \frac{U_{ij}}{\sum_i U_{ij}} 10^6 \\ E_{kj} &= \frac{C_{kj} - \bar{M}_k(C_{kj})}{\hat{V}_k(C_{kj})} \\ S_j &= \bar{M}_k(F_{kj}) \end{aligned} \tag{3}$$

where U is the UMI count for gene i in cell j. $\bar{M}$ is the the mean, and $\hat{V}$ the variance along the indexed dimension. F is calculated by capping per gene k scaled expression vector E_{kj} at the 0.01 - and 0.99 percentile. Genes k are the genes selected for score calculation. This approach does not weight genes with generally greater UMI counts, than low-UMI-counts genes, as done in La Manno *et al.*.

6.13 Cell subset integration

Due to the mixing of all ventricular zone associated cells in the low dimensional embedding no cluster based selection could be performed. Though, differential gene expression within clusters between developmental stages allowed a marker based selection of the presumed Purkinje and interneuron precursors, respectively: Purkinje associated VZ in mouse were picked from E12.5 and E13.5. Interneuron progenitors were chosen from VZ neuroblasts older than E13.5. In human, $\mathrm{VZ}_{\mathrm{neuroblast1}}$ cells from Carnegie stages 18-19, $\mathrm{VZ}_{\mathrm{neuroblast2}}$ cells from Carnegie stages 18-22 and $\mathrm{VZ}_{\mathrm{neuroblast3}}$ cells from Carnegie stages 19-22 were assigned to Pukinje lineage and older than CS22 to inteneurons. In opossum all VZ cells expressing *PAX2* and/or *SLC16A5* were labelled as interneurons in addition to VZ cells older than P14. Purkinje progenitors were VZ cells, *LMX1A* and *LMX1B* negative and from stages P4 and P5.

Subsets of cells were integrated as described above, using LIGER for batch correction. The number of components per non-negative matrix factorization was chosen as follows: Purkinje cells: mouse ($k = 70$), human ($k = 50$), and opossum ($k = 70$). Granule cells: mouse ($k = 30$), human ($k = 40$), and opossum ($k = 40$).

6.14 Cross-species correlation

Subtypes matches for Purkinje cells and interneurons were achieved by Spearmans correlation analysis. For interneurons, cells originating in human deep nuclei (n = 57) were removed prior to the correlation coefficient determination. The geneset was defined by shared highly variable one-to-one

ortholgs. The number of used orthologs in cross-species correlation of Purkinje subtypes was 107, of interneuron subtypes was 198.

6.15 Comparison to adult mouse data by Kozareva *et al.*

Adult mouse data was collected as UMI count matrices from Kozareva *et al.*. First, I subsetted the mouse data to Purkinje or granule cells associated cells (as described above) and did the same for the published data. Overdispersed genes were called independently and intersected between both studies. Subtype pseudobulks were generated as previously described and normalized matrices were used for Spearmans correlation analysis.

6.16 Principal component analysis

Principal component analysis (PCA) was applied to investigate generally shared expression profiles between the studied species. First, groups defined by biological replicate, stage and cell type were combined to pseudobulks, as described above, using only three-way 1:1 orthologs. Only such pseudobulks were kept, that contained at least 150 cells. Genes which were not expressed in at least 10% of any pseudobulk, and did not show variability ($\hat{V}(CPM) > 0$) were removed to improve the signal-to-noise ratio. Expression vectors per species and gene were median-centered. The matrix was combined and PCA was conducted using the `prcomp_irlba` function of the irlba [132] package (v2.3.3).

6.17 Conserved marker gene calling

Only cell states and stages which were shared between species were used for downstream processing. To overcome differences in sampling per species and cell state, cell states were randomly sampled to 1,000 cells. If the number of cells was lower than 1,000 cells, the present cells were randomly upsampled to 1,000. Per species, marker genes were identified for each cell state, as described previously. Marker genes were then filtered to show an enrichment of at least two fold, a false discovery rate of < 0.01 and a percentage of expressing cells within the cell state of at least 10%. The intersect of the genes, passing this filter in all three species were ranked according to their distance to the origin of the Cartesian coordinate system in three-dimensional TF-IDF (T) space (equation 4).

$$S_j = \sqrt{T_{j,human}^2 + T_{j,mouse}^2 + T_{j,opossum}^2} \tag{4}$$

6.18 Pseudotime calling

To investigate genes that are regulated along Purkinje cell or granule cell development, I first extracted cells of both lineages from the dataset (as explained previously). Genes were subsetted to the three-way 1:1-orthologs and data was integrated per cell type using the Harmony [76] pipeline. I decided to call the pseudotemporal vector on the integrated dataset to prevent species-specific variations in assignment and to remove the need of a *post-hoc* alignment of the pseudotime using dynamic time warping. The starting cell for the integrated pseudotime calling was chosen based on a learned UMAP embedding, which was trained on the Harmony corrected components. The pseudotime vector was determined using the SCANPY implementation of diffusion pseudotime [114].

6.19 Expression trajectories along pseudotime vectors

To capture gene expression changes along the determined pseudotemporal ordering in Purkinje and granule cells, I binned the pseudotime vector into ten equally sized bins. Per bin, UMI values were averaged across biological replicates, if the pseudobulk consisted of at least 50 cells. Due to the non-linear nature of some expected trajectories, instead of calculating correlation coefficients, I filtered for highly variable genes, by applying the aforementioned strategy to the binned and averaged UMI counts ($\alpha = 1$). Only genes which were called as highly variable in all three species were subjected to the following algorithm: CPM values were calculated per species and expression values were scaled. Next, each ortholog was assigned to the specific species, the genes signal was generated from and the resulting three matrices were combined that all members of a three-way ortholog group were present in the final matrix[8]. To identify major trajectory patterns, fuzzy clusters were called using the Mfuzz package [133] (v2.44.0), allowing for eight trajectory clusters (fuzzy parameter set to 1.2). Cluster members were accepted as confident when the membership value scored higher than 0.5. Per feature, the center of mass was calculated (custom function). To rank fuzzy clusters, all center of mass values were averaged across confident members and results were sorted increasingly. If a given gene scored lower than 0.5 in any species, the orthologs were removed from any further analysis. Similarity between pairwise comparisons of orthologs was accomplished by calculating the agreement between cluster memberships p as shown in equation 5

$$p(x, y) = \sum_{i=1}^{k} m_{xi} m_{yi} \tag{5}$$

where the agreement between species x and y for orthologous gene i is determined using the cluster membership vector m per ortholog.

Classification was applied according to the following rules:

[8] This means, that each gene was present three times, once per species.

$$class_i = \begin{cases} \text{x specific} & p(x,y) < 0.05 \wedge p(x,z) < 0.05 \wedge p(y,z) > 0.5 \wedge C_y = C_z \\ \text{conserved} & C_x = C_y = C_z \wedge p(x,y) > 0.5 \wedge p(y,z) > 0.5 \\ \text{diverse} & p(x,y) < 0.05 \wedge p(y,z) < 0.05 \wedge p(x,z) < 0.05 \\ \text{intermediate} & \text{otherwise} \end{cases} \tag{6}$$

where C represents the cluster call for species x, y, or z ortholog, according to maximum membership assignment. As additional measure, dynamic timewarp distance was measured using the dtw package in R [131]. Per ortholog group the maximum and minimum pairwise distance was evaluated. Patterns of change were visualized using alluvial plots (ggalluvial [134], v0.12.3)

6.20 Gain and loss classification

The approach to find candidate genes which gained or lost expression in a species-specific fashion in a specific cell type can be summarised as follows: I focused on the following cell types: astroglia, GABAergic deep nuclei neurons, glutamatergic deep nuclei neurons, granule cells, interneurons, oligodendrocytes, Purkinje cells, unipolar brush cells. Only exonic UMI counts were considered to reduce the effect of differential abundance of intronic poly-A stretches. As previously described, pseudobulk were generated per cell type annotation, species and biological replicate and kept if more than 50 cells were grouped. Expression was normalized to CPM and maximum expression per cell type and species determined. A cutoff of 50 CPM was chosen to differentiate between confidently expressed genes and lowly expressed genes. Additionally, genes were removed, which did not show expression in the single nuclei dataset, but exceeded 5 RPKM in bulk RNA-seq of the cerebellum [7]. Each cell types maximum expression was contrasted against the maximum expression within the dataset of the selected cell types within the same species. This percent of maximum expression was used to identify genes which were expressed above background, using a cutoff of 0.3.

The classification strategy is described in detail in the results (section 3.9) and equation 2.

6.21 Cell type specificity

Inspired by Yanai *et al.* [98], I applied specificity τ to assess cell type specificity of gene expression, rather tissue specificity. Specificity τ was calculated as shown in equation 7.

$$\begin{aligned} \tau &= \frac{\sum_{i=1}^{n} 1 - \hat{x}_i}{n-1} \\ \hat{x}_i &= \frac{x_i}{max(x)} \end{aligned} \tag{7}$$

where gene x normalized expression is normalized ($\hat{x}$) to the maximal detected expression within a species across all cell types and then evaluated against the total number of all detected cell types

n. This scaling represents broad expression for values close to 0 and absolute cell type specificity at 1.

6.22 Gene onthology enrichment analysis

Gene onthology annotations were collected from ENSEMBL version 91 for the whole mouse genome and subsetted for three-way orthologs between human, mouse and opossum. Enrichment p value was calculated applying the observed counts to the `pbinom` function in R. I accounted for multiple testing by applying the Benjamini-Hochberg method to all observed p-values using th `p-adjust` function in R.

6.23 Adult bulk gene expression determination

RNA-seq data was retrieved from Brawand *et al.* [100] for chicken and nine mammalian species adult brain and cerebellum. Data was aligned against the ENSEMBL version 91 genomes and annotations. As described in Wang *et al.* [125], expression values were determined on FPKM scale. Per gene, the longest protein coding isoform was considered which perfectly align between species.

6.24 Disease gene annotation retrieval

Loss-of-function observed/expected upper bound fraction scores (LOEUF) [93] were retrieved from the Genome Aggregation Database (gnomAD). Genes are ranked according to their tolerance of *in vivo* loss-of-function alterations, judged on human genome and exome sequecing results.

In vivo essentiality scores by Bartha *et al.* [6], aggregating various essentiality scores (RVIS, pLI, Phi, missense Z-score, LoFtool and s_{het}) were retreived from the original publication. This measure is based on human exome and genome sequencing data.

In vitro essentiality scores were retrieved from the same source [6]. This score is based on *in-vitro* CRISPR-Cas9 inactivation screens.

The Human Gene Mutation Database (HGMD, PRO 17.1) [135] provided the human inherited disease gene list[9]. The genes were subsetted based on the Unified Medical Language System (UMLS) to genes which are linked to the following disease types (high level annotation): 'Nervous system' and 'Psychiatric'. Depending on the high level disease type 'Development', the filtered genes were grouped into development associated genes, or non-associated genes.

Cerebellum-linked disease lists were collected from Aldinger *et al.* [95] and Gröbner *et al.* [136]. Aldinger *et al.* lists were the following: (I) Genes associated with cerebellar malformations, Dandy-Walker malformations and hypoplasia. (II) Joubert syndrome list, (III) autism spectrum list, (III) intellectual disability lists, and (IV) Spinocerebellar ataxia list. From Gröbner *et al.* [136] the

[9] Which is manually currated.

pediatric cancer driver gene list was retrieved, including medulloblastoma, ependymoma, pilocytic astroctoma, and pleomorphic xanthoastrocytoma.

Enrichments of discrete associations were calculated by applying the `pbinom` (R) to the counts per category, using the list of highly variable genes as the possible gene universe. Scores which are on continuous scale were investigated using permutation test (n = 10,000), accounting for intolerance directionality by adjusting the alternative hypothesis accordingly.

6.25 General toolset

All analyses, if not stated ortherwise were conducted with R (v3.6). The following R packages were used for analyses and plotting: tidyverse [137] (v1.3), SingleCellExperiment [138] (v1.6), LIGER [139] (v0.4.6), rliger [77] (v1.0), batchelor [75] (v1.0.1), pheatmap (v1.0.12), ggplot2 [140] (v3.3.2). Python (v3.6) was used for the following packages: SCANPY [74] (v1.5.1) and htseq [141] (v0.13.5).

7 Acknowledgments / Danksagungen

Diese Arbeit war ein hartes Stück Arbeit, welches fast sechs Jahre meines Lebens brauchte. Ohne die Unterstützung der folgenden Menschen wäre sie nicht möglich gewesen.

Zuerst möchte ich mich bei Prof. Henrik Kaessmann für die Möglichkeit bedanken in seinem Labor diese Arbeit anzufertigen. Er vertraute meinen Fähigkeiten, half mir bei Problemen und war in all der Zeit immer verständnisvoll und auch für einen Spaß zu haben.

Dann gilt mein ganzer Dank Dr. Mari Sepp. Ohne sie wäre diese Arbeit wortwörtlich nicht möglich gewesen. Ihre alltägliche Unterstützung, Hilfe bei Interpretationen und Ideen die Daten aus einem anderen Blickwinkel zu betrachten, sind unermesslich. Ich denke wir waren ein tolles Team. Auch danke ich ihr für die Zeit, die sie sich nahm, diese Arbeit kritisch zu lesen. Sie gab mir viel wertvolles Feedback.

Natürlich möchte ich mich auch bei unserem gesamten Labor bedanken. Ich könnte hier nun jeden einzelnen aufzählen, da alle eine tolle und produktive Umgebung geschaffen haben. An dieser Stelle möchte ich mich aber ganz beßonders bei Ioannis Sarropoulos, Florent Murat, Nils Trost, Celine Schneider und Julia Schmitt bedanken.

Kornelia Mack und Kathrin Hall halfen mir bei allen administrativen Aufgaben und waren echte Meister, Probleme zu lösen und unterstützen mich bei allen administrativen Aufgaben.

Mein Dank gilt auch den TAC Mitgliedern, Prof. Georg Stoecklin und Prof. Detlev Arendt. Beide gaben immer sehr wichtigen Input während unseren Meetings und hatten immer ein offenes Ohr.

Auch bei Prof. Ana Martin-Villalba möchte ich mich bedanken. In ihrem Labor machte ich mein erstes eigenständiges Laborpraktikum während meines Bachelors. Ich lernte viel und es gab mir den Mut und das Selbstvertrauen, das man in der Forschung braucht.

Des Weiteren gilt mein Dank Prof. Christine Clayton. Sie war die Erste, die mir die Möglichkeit gab, bioinformatisch an einem unbglaublich interessanten Projekt zu arbeiten, obwohl ich mit nur wenig Vorerfahrung bei ihr anfing. Selbst in den Jahren nach meiner Masterarbeit bei ihr, publizierte sie weiterhin Analysen, an denen ich beteiligt war und nahm mich immer in die Autorenliste auf, egal wie klein mein Zutun war. Christine gab meiner wissenschaftlichen Karriere einen echten Blitzstart.

Meinen Eltern danke ich für ihre immerwährende Unterstützung und Liebe. Egal welche Steine in meinem Leben vor mich gelegt wurden, ich wusste immer, dass ich mich auf sie verlassen kann - Egal was es kostete und wie viel Zeit es in Anspruch nahm. Auch meinem Bruder, Steven, der gleichfalls mein bester Freund ist, gebührt mein Dank für Alles, was er für mich getan hat - wir sind wie Pech und Schwefel. Es ist nicht zu bemessen, wie glücklich ich bin eine solche Familie zu haben.

Und natürlich danke ich meiner besseren Hälfte, Boyana. Ich bin so glücklich, dass ich vor über

elf Jahren den Zug verpasste und sie in der Zoologie in Heidelberg kennen lernen durfte. Boyana glaubte immer an mich und dass wir alles zusammen schaffen können. Sie gibt mir den Halt, den ich brauche, weil ich weiß, dass ich mich auf sie zu 100% verlassen kann. Ihr habe ich es zu verdanken, dass ich überhaupt den Mut fand, bioinformatisch zu arbeiten. Sie ist meine große Liebe, beste Freundin und Ratgeberin. Ich freue mich auf die kommenden Jahrzente mit dir und bin gespannt, was unsere Zukunft uns noch bringen wird.

8 Abbreviations

8.1 General

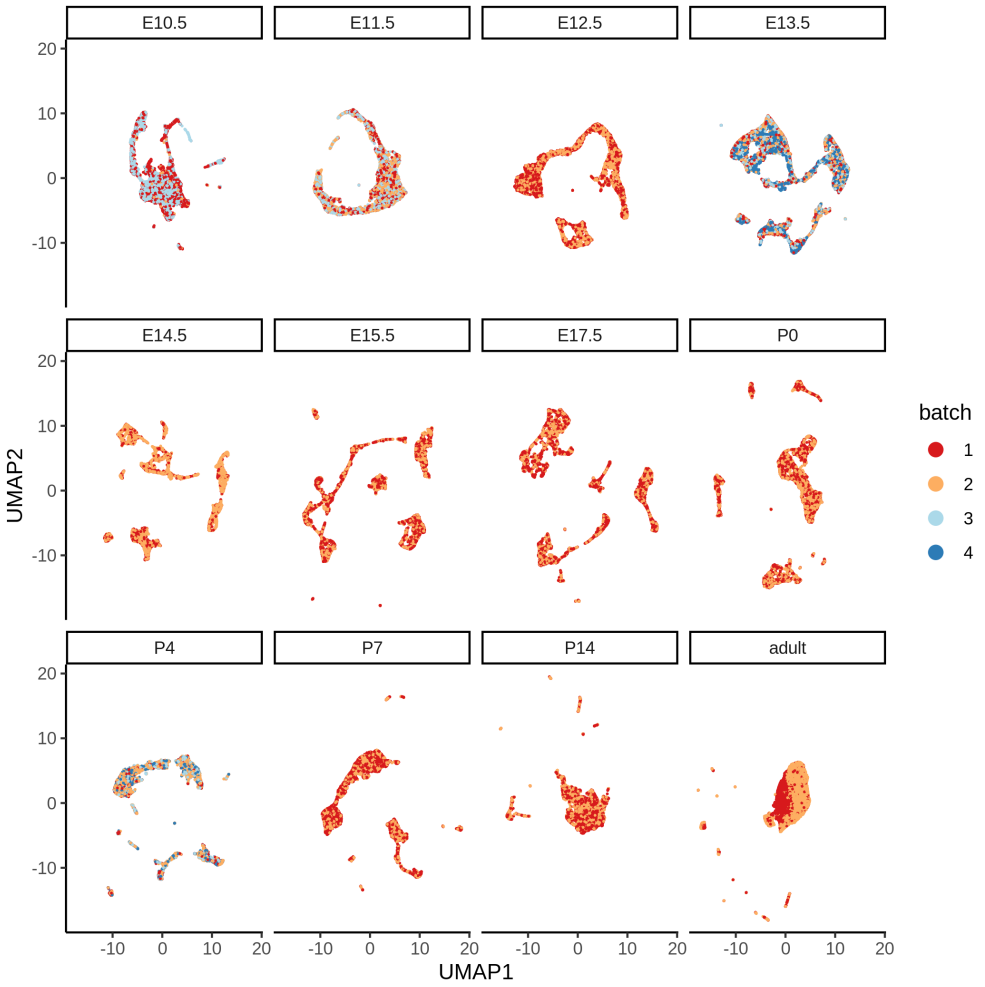
- **E10.5** (and similar): embryonic day 10.5
- **HVG**: highly variable genes, informative genes
- **GO**: gene onthology
- **P0** (and similar): Postnatal day 0 (birth)
- **PC**: principal component
- **PCA**: principal component analysis
- **RNA-seq**: RNA sequencing, transcriptomic profiling method
- **snRNA-seq**: single nucleus RNA-seq
- **TF-IDF**: term-frequency inversed document frequency
- **UMAP**: Uniform manifold approximation and projection
- **UMI**: unique molecular identifier
- **wpc**: weeks post conception

8.2 Cell types / cerebellar structures

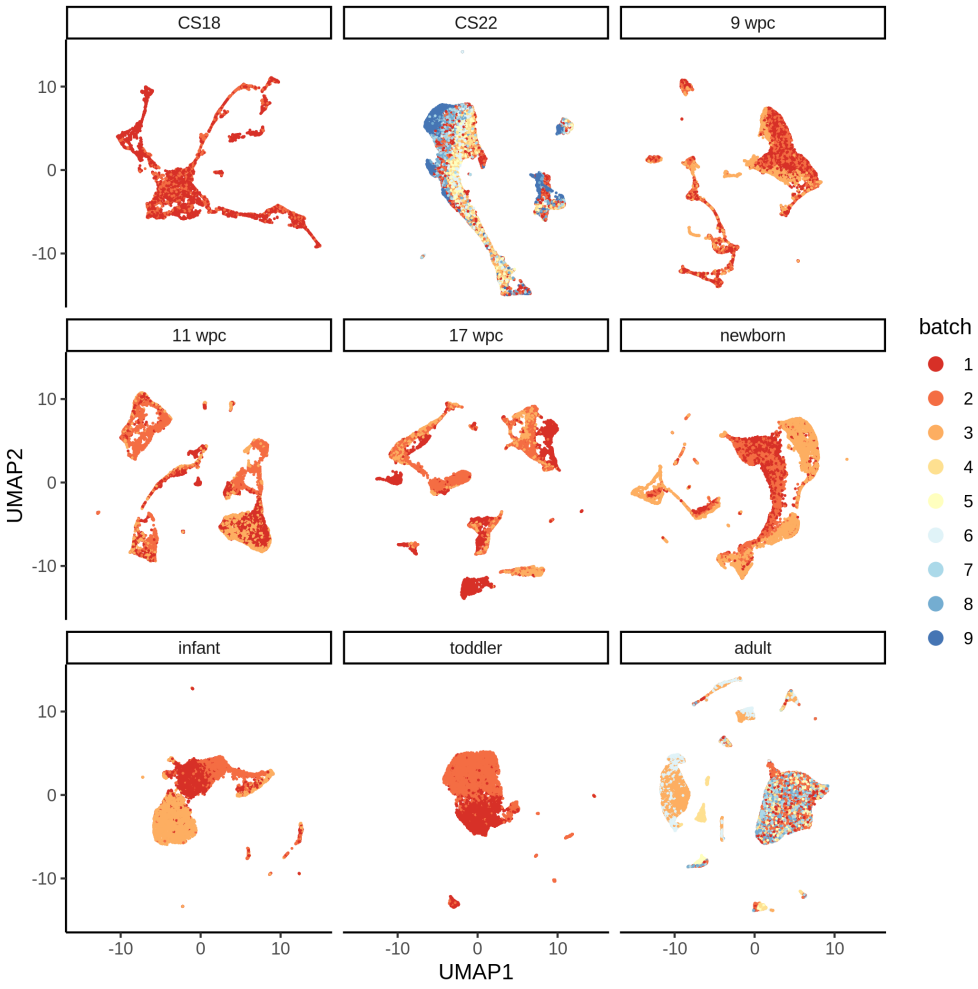
- **DN**: deep nuclei
- **EGL**: external granule cell layer
- **GABA**: gamma-aminobutyric acid
- **GC**: granule cells
- **IGL**: internal granule cell layer
- **ML**: molecular layer
- **MB**: midbrain
- **NTZ**: neuronal transitory zone
- **PC**: Purkinje cells
- **PL**: Purkinje layer

- **RL**: rhombic lip
- **UBC**: unipolar brush cells
- **VZ**: ventricular zone

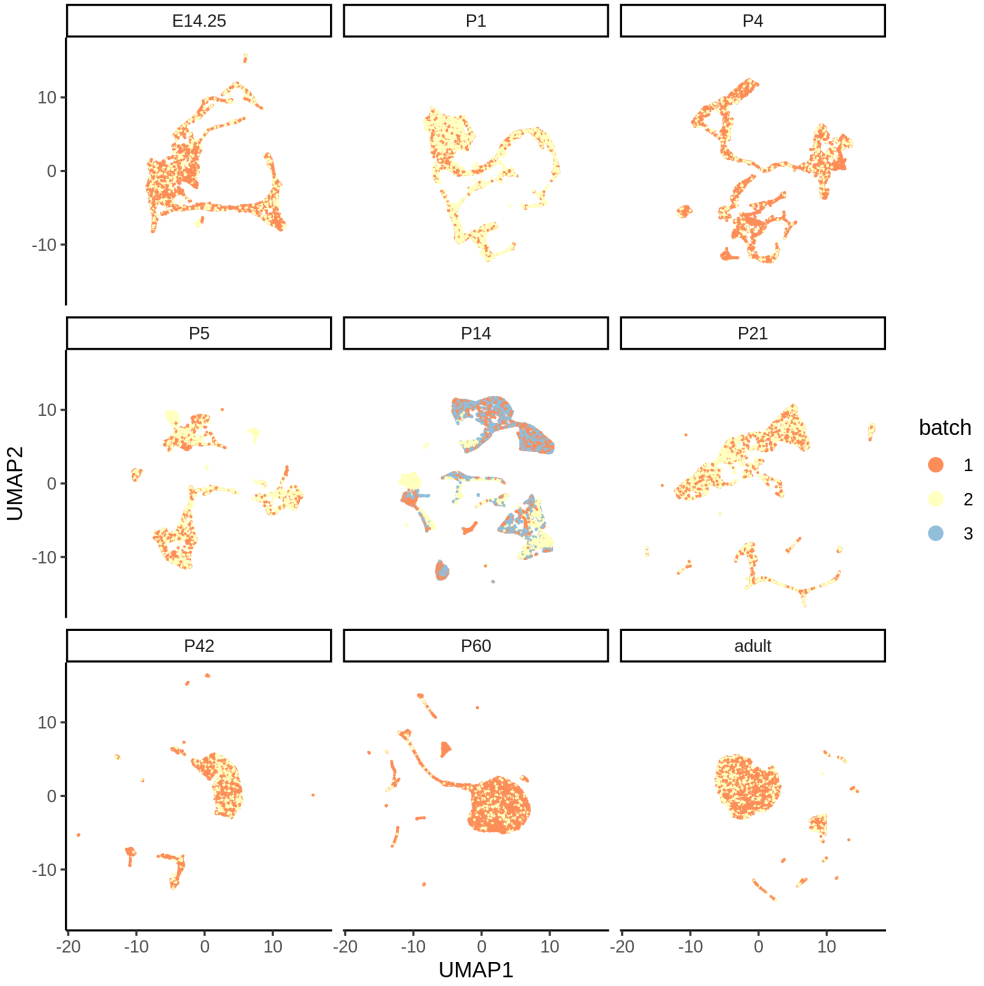
9 Supplementary figures



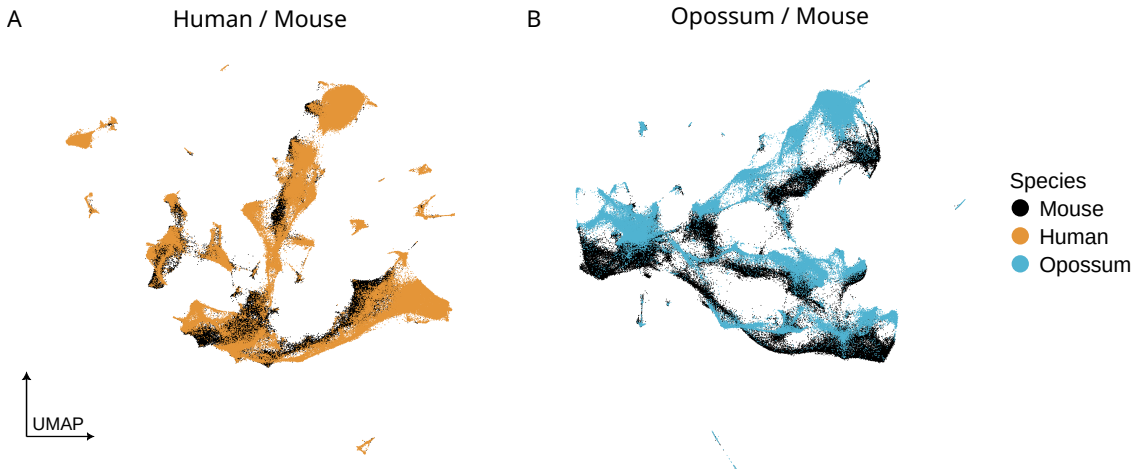
Supplementary Figure S1: Mouse UMAP embeddings per stage Embeddings were generated from LIGER corrected batch integrations per stage. Colors represent individual libraries.



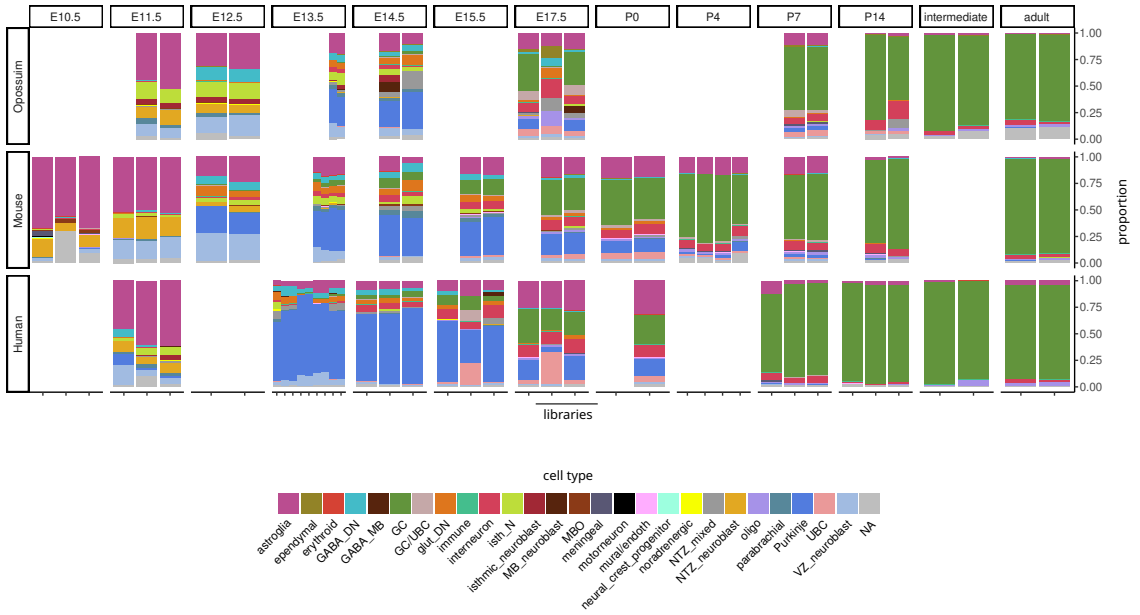
Supplementary Figure S2: Human UMAP embeddings per stage Embeddings were generated from LIGER corrected batch integrations per stage. Colors represent individual libraries.



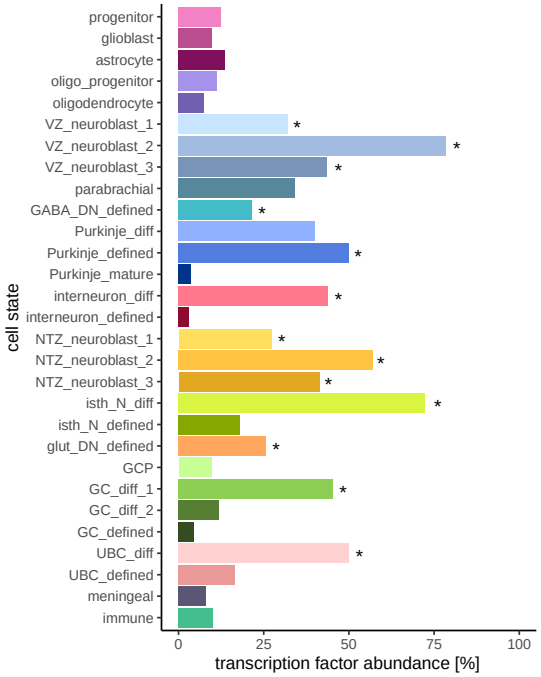
Supplementary Figure S3: Opossum UMAP embeddings per stage Embeddings were generated from LIGER corrected batch integrations per stage. Colors represent individual libraries.



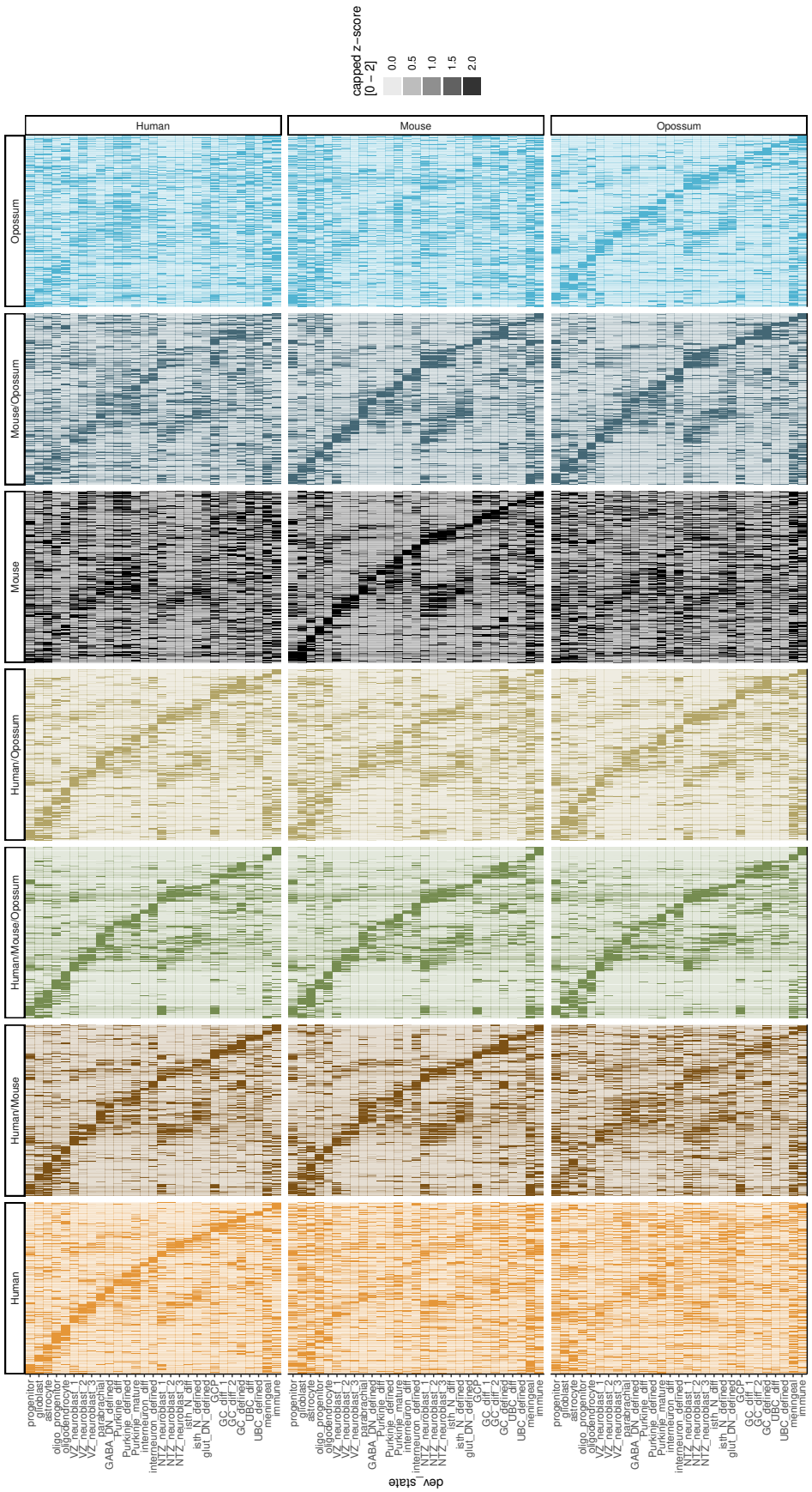
Supplementary Figure S4: UMAPs of pairwise species integration A: UMAP generated from LIGER and MNN corrected embedding of all human and mouse data. **A**: UMAP generated from LIGER and MNN corrected embedding of all opossum and mouse data.



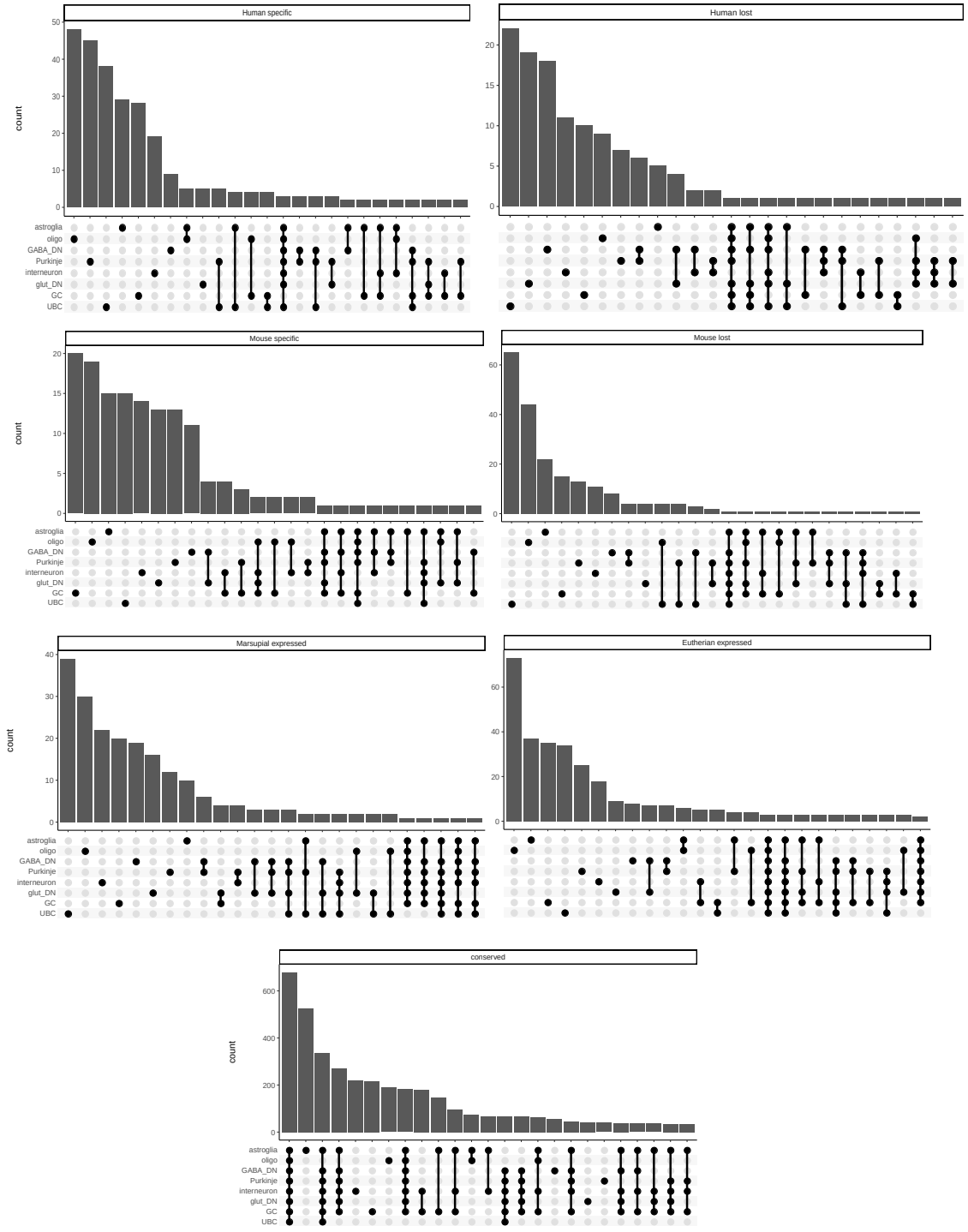
Supplementary Figure S5: Cell type abundances per batch Per batch and species for aligned developmental timepoints (horizontal facets), the abundances for each cell type was determined. Each stacked bar represents a single single nucleus library.



Supplementary Figure S6: Transcription factor proportion in conserved cell state markers Significant enrichments of transcription factors among the depicted groups are indicated with an asterisk.



Supplementary Figure S7: Marker gene heatmap across all groups Same depiction as in figure 11, but containing all possible combinations of shared marker genes across the three species.



Supplementary Figure S8: Presence absence calls overlaps Classes of presence and absence calls are shown with the number of shared classifications across the observed cell types.

10 List of figures

11 List of tables

12 References

1. Buckner, R. L. The Cerebellum and Cognitive Function: 25 Years of Insight from Anatomy and Neuroimaging. *Neuron* **80,** 807–815. ISSN: 0896-6273 (Oct. 2013).
2. Butts, T., Green, M. J. & Wingate, R. J. T. Development of the Cerebellum: Simple Steps to Make a 'Little Brain'. *Development* **141,** 4031–4041. ISSN: 0950-1991 (Nov. 2014).
3. Butts, T., Modrell, M. S., Baker, C. V. & Wingate, R. J. The Evolution of the Vertebrate Cerebellum: Absence of a Proliferative External Granule Layer in a Non-Teleost Ray-Finned Fish. *Evolution & Development* **16,** 92–100. ISSN: 1525-142X (2014).
4. Leto, K., Arancillo, M., Becker, E. B. E., Buffo, A., *et al.* Consensus Paper: Cerebellar Development. *The Cerebellum* **15,** 789–828. ISSN: 1473-4230 (Dec. 2016).
5. Haldipur, P., Aldinger, K. A., Bernardo, S., Deng, M., *et al.* Spatiotemporal Expansion of Primary Progenitor Zones in the Developing Human Cerebellum. *Science* **366,** 454–460 (Oct. 2019).
6. Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human Gene Essentiality. *Nature Reviews Genetics* **19,** 51–62. ISSN: 1471-0064 (Jan. 2018).
7. Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., *et al.* Gene Expression across Mammalian Organ Development. *Nature* **571,** 505–509. ISSN: 1476-4687 (July 2019).
8. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134,** 25–36. ISSN: 0092-8674 (July 2008).
9. Kozareva, V., Martin, C., Osorno, T., Rudolph, S., Guo, C., Vanderburg, C., Nadaf, N., Regev, A., Regehr, W. G. & Macosko, E. A Transcriptomic Atlas of Mouse Cerebellar Cortex Comprehensively Defines Cell Types. *Nature* **598,** 214–219. ISSN: 1476-4687 (Oct. 2021).
10. Sepp, M., Leiss, K., Sarropoulos, I., Murat, F., *et al.* Cellular Development and Evolution of the Mammalian Cerebellum. *bioRxiv,* 2021.12.20.473443 (Dec. 2021).
11. Breasted, J. H. THE EDWIN SMITH SURGICAL PAPYRUS. *THE UNIVERSITY OF CHICAGO ORIENTAL INSTITUTE PUBLICATIONS* **1,** 634 (1980).
12. Longrigg, J. Anatomy in Alexandria in the Third Century B.C. *British Journal for the History of Science* **21,** 455–488. ISSN: 0007-0874 (Dec. 1988).
13. Clarke, E. & O'Malley, C. D. *The Human Brain and Spinal Cord: A Historical Study Illustrated by Writings from Antiquity to the Twentieth Century* ISBN: 978-0-930405-25-0 (Norman Publishing, 1996).

14. Stoodley, C. J. & Schmahmann, J. D. Evidence for Topographic Organization in the Cerebellum of Motor Control versus Cognitive and Affective Processing. *Cortex; a journal devoted to the study of the nervous system and behavior* **46,** 831–844. ISSN: 0010-9452 (2010).

15. Flourens, P. *Recherches Experimentales Sur Les Propietes et Les Functions Du Systeme Nerveux* (Crevot, 1824).

16. Rolando, L. *Saggio sopra la vera struttura del cervello dell'uomo e degl'animali e sopra le funzioni del sistema nervoso di Luigi Rolando ..* (nella stamperia da S.S.R.M privilegiata, 1809).

17. Sathyanesan, A., Zhou, J., Scafidi, J., Heck, D. H., Sillitoe, R. V. & Gallo, V. Emerging Connections between Cerebellar Development, Behaviour and Complex Brain Disorders. *Nature Reviews Neuroscience* **20,** 298–313. ISSN: 1471-0048 (May 2019).

18. Ostrom, Q. T., Gittleman, H., Xu, J., Kromer, C., Wolinsky, Y., Kruchko, C. & Barnholtz-Sloan, J. S. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2009–2013. *Neuro-Oncology* **18,** v1–v75. ISSN: 1522-8517 (Oct. 2016).

19. Guerreiro Stucklin, A. S. & Grotzer, M. A. in *Handbook of Clinical Neurology* (eds Manto, M. & Huisman, T. A. G. M.) 289–299 (Elsevier, Jan. 2018).

20. Cajal, R. Histologie Du Systeme Nerveux de l'Homme et Des Vertebres. Grand Sympathique. *Paris Maloine* **2,** 891–942 (1911).

21. White, J. J. & Sillitoe, R. V. Development of the Cerebellum: From Gene Expression Patterns to Circuit Maps. *WIREs Developmental Biology* **2,** 149–164. ISSN: 1759-7692 (2013).

22. Palay, S. L. & Chan-Palay, V. *Cerebellar Cortex: Cytology and Organization* ISBN: 978-3-642-65581-4 (Springer Science & Business Media, Dec. 2012).

23. Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C. & Yeo, B. T. T. The Organization of the Human Cerebellum Estimated by Intrinsic Functional Connectivity. *Journal of Neurophysiology* **106,** 2322–2345. ISSN: 0022-3077 (Nov. 2011).

24. Leiner, H. C. Solving the Mystery of the Human Cerebellum. *Neuropsychology Review* **20,** 229–235. ISSN: 1573-6660 (Sept. 2010).

25. Brochu, G., Maler, L. & Hawkes, R. Zebrin II: A Polypeptide Antigen Expressed Selectively by Purkinje Cells Reveals Compartments in Rat and Fish Cerebellum. *Journal of Comparative Neurology* **291,** 538–552 (1990).

26. Ahn, A. H., Dziennis, S., Hawkes, R. & Herrup, K. The Cloning of Zebrin II Reveals Its Identity with Aldolase C. *Development* **120,** 2081–2090 (1994).

27. Apps, R. & Hawkes, R. Cerebellar Cortical Organization: A One-Map Hypothesis. *Nature Reviews Neuroscience* **10,** 670–681 (2009).

28. Foucher, I., Mione, M., Simeone, A., Acampora, D., Bally-Cuif, L. & Houart, C. Differentiation of Cerebellar Cell Identities in Absence of Fgf Signalling in Zebrafish Otx Morphants (2006).

29. Crossley, P. H., Martinez, S. & Martin, G. R. Midbrain Development Induced by FGF8 in the Chick Embryo. *Nature* **380,** 66–68 (1996).

30. Martinez, S., Crossley, P. H., Cobos, I., Rubenstein, J. L. & Martin, G. R. FGF8 Induces Formation of an Ectopic Isthmic Organizer and Isthmocerebellar Development via a Repressive Effect on Otx2 Expression. *Development* **126,** 1189–1200 (1999).

31. Sato, T. & Joyner, A. L. The Duration of Fgf8 Isthmic Organizer Expression Is Key to Patterning Different Tectal-Isthmo-Cerebellum Structures. *Development* **136,** 3617–3626 (2009).

32. Liu, A. & Joyner, A. L. Early Anterior/Posterior Patterning. *Annu. Rev. Neurosci* **24,** 869–96 (2001).

33. Zervas, M., Blaess, S. & Joyner, A. L. Classical Embryological Studies and Modern Genetic Analysis of Midbrain and Cerebellum Development. *Current topics in developmental biology* **69,** 101–138 (2005).

34. Wingate, R. J. & Hatten, M. E. The Role of the Rhombic Lip in Avian Cerebellum Development. *Development* **126,** 4395–4404 (1999).

35. Anthony, T. E., Klein, C., Fishell, G. & Heintz, N. Radial Glia Serve as Neuronal Progenitors in All Regions of the Central Nervous System. *Neuron* **41,** 881–890 (2004).

36. Hoshino, M., Nakamura, S., Mori, K., Kawauchi, T., *et al.* Ptf1a, a bHLH Transcriptional Gene, Defines GABAergic Neuronal Fates in Cerebellum. *Neuron* **47,** 201–213. ISSN: 0896-6273 (July 2005).

37. Machold, R. & Fishell, G. Math1 Is Expressed in Temporally Discrete Pools of Cerebellar Rhombic-Lip Neural Progenitors. *Neuron* **48,** 17–24 (2005).

38. Wang, V. Y., Rose, M. F. & Zoghbi, H. Y. Math1 Expression Redefines the Rhombic Lip Derivatives and Reveals Novel Lineages within the Brainstem and Cerebellum. *Neuron* **48,** 31–43 (2005).

39. Pascual, M., Abasolo, I., Mingorance-Le Meur, A., Martínez, A., Del Rio, J. A., Wright, C. V., Real, F. X. & Soriano, E. Cerebellar GABAergic Progenitors Adopt an External Granule Cell-like Phenotype in the Absence of Ptf1a Transcription Factor Expression. *Proceedings of the National Academy of Sciences* **104,** 5193–5198 (2007).

40. Englund, C., Kowalczyk, T., Daza, R. A., Dagan, A., Lau, C., Rose, M. F. & Hevner, R. F. Unipolar Brush Cells of the Cerebellum Are Produced in the Rhombic Lip and Migrate through Developing White Matter. *Journal of Neuroscience* **26,** 9184–9195 (2006).

41. Hashimoto, M. & Mikoshiba, K. Mediolateral Compartmentalization of the Cerebellum Is Determined on the "Birth Date" of Purkinje Cells. *Journal of Neuroscience* **23,** 11342–11351. ISSN: 0270-6474, 1529-2401 (Dec. 2003).

42. Namba, K., Sugihara, I. & Hashimoto, M. Close Correlation between the Birth Date of Purkinje Cells and the Longitudinal Compartmentalization of the Mouse Adult Cerebellum. *Journal of Comparative Neurology* **519,** 2594–2614. ISSN: 1096-9861 (2011).

43. Hevner, R. F. in *Reelin Glycoprotein* 141–158 (Springer, 2008).

44. Miyata, T., Ono, Y., Okamoto, M., Masaoka, M., Sakakibara, A., Kawaguchi, A., Hashimoto, M. & Ogawa, M. Migration, Early Axonogenesis, and Reelin-dependent Layer-Forming Behavior of Early/Posterior-Born Purkinje Cells in the Developing Mouse Lateral Cerebellum. *Neural development* **5,** 1–22 (2010).

45. Sillitoe, R. V., Gopal, N. & Joyner, A. L. Embryonic Origins of ZebrinII Parasagittal Stripes and Establishment of Topographic Purkinje Cell Projections. *Neuroscience* **162,** 574–588 (2009).

46. Chung, S. .-., Marzban, H., Croci, L., Consalez, G. G. & Hawkes, R. Purkinje Cell Subtype Specification in the Cerebellar Cortex: Early B-cell Factor 2 Acts to Repress the Zebrin II-positive Purkinje Cell Phenotype. *Neuroscience* **153,** 721–732. ISSN: 0306-4522 (May 2008).

47. Croci, L., Chung, S.-H., Masserdotti, G., Gianola, S., Bizzoca, A., Gennarini, G., Corradi, A., Rossi, F., Hawkes, R. & Consalez, G. G. A Key Role for the HLH Transcription Factor EBF2COE2, O/E-3 in Purkinje Neuron Migration and Cerebellar Cortical Topography (2006).

48. Edmondson, J. C. & Hatten, M. E. Glial-Guided Granule Neuron Migration in Vitro: A High-Resolution Time-Lapse Video Microscopic Study. *Journal of Neuroscience* **7,** 1928–1934 (1987).

49. ten Donkelaar, H. J., Lammens, M., Wesseling, P., Thijssen, H. O. & Renier, W. O. Development and Developmental Disorders of the Human Cerebellum. *Journal of neurology* **250,** 1025–1036 (2003).

50. Víg, J., Takács, J., Abraham, H., Kovács, G. G. & Hámori, J. Calretinin-Immunoreactive Unipolar Brush Cells in the Developing Human Cerebellum. *International journal of developmental neuroscience* **23,** 723–729 (2005).

51. Miale, I. L. & Sidman, R. L. An Autoradiographic Analysis of Histogenesis in the Mouse Cerebellum. *Experimental Neurology* **4,** 277–296. ISSN: 0014-4886 (Oct. 1961).

52. Maricich, S. M. & Herrup, K. Pax-2 Expression Defines a Subset of GABAergic Interneurons and Their Precursors in the Developing Murine Cerebellum. *Journal of neurobiology* **41,** 281–294 (1999).

53. Wolf, S. Development of the Cerebellar System in Relation to Its Evolution, Structure, and Functions. *Integrative Physiological and Behavioral Science* **35,** 71–71. ISSN: 1053881X (Jan. 2000).

54. Weisheit, G., Gliem, M., Endl, E., Pfeffer, P. L., Busslinger, M. & Schilling, K. Postnatal Development of the Murine Cerebellar Cortex: Formation and Early Dispersal of Basket, Stellate and Golgi Neurons. *European Journal of Neuroscience* **24,** 466–478 (2006).

55. Wechsler-Reya, R. J. & Scott, M. P. Control of Neuronal Precursor Proliferation in the Cerebellum by Sonic Hedgehog. *Neuron* **22,** 103–114. ISSN: 0896-6273 (Jan. 1999).

56. Herculano-Houzel, S., Manger, P. R. & Kaas, J. H. Brain Scaling in Mammalian Evolution as a Consequence of Concerted and Mosaic Changes in Numbers of Neurons and Average Neuronal Cell Size. *Frontiers in Neuroanatomy* **8.** ISSN: 1662-5129 (2014).

57. Barton, R. A. Embodied Cognitive Evolution and the Cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367,** 2097–2107 (2012).

58. Libé-Philippot, B. & Vanderhaeghen, P. Cellular and Molecular Mechanisms Linking Human Cortical Development and Evolution. *Annual review of genetics,* 555–581 (2021).

59. Pinson, A. & Huttner, W. B. Neocortex Expansion in Development and Evolution—from Genes to Progenitor Cell Biology. *Current opinion in cell biology* **73,** 9–18 (2021).

60. Barton, R. A. & Venditti, C. Rapid Evolution of the Cerebellum in Humans and Other Great Apes. *Current Biology* **24,** 2440–2444. ISSN: 0960-9822 (Oct. 2014).

61. Neubauer, S., Hublin, J.-J. & Gunz, P. The Evolution of Modern Human Brain Shape. *Science advances* **4,** eaao5961 (2018).

62. Kebschull, J. M., Richman, E. B., Ringach, N., Friedmann, D., *et al.* Cerebellar Nuclei Evolved by Repeatedly Duplicating a Conserved Cell-Type Set. *Science* **370,** eabd5059 (Dec. 2020).

63. Skinner, M. K., Gurerrero-Bosagna, C., Haque, M. M., Nilsson, E. E., Koop, J. A., Knutie, S. A. & Clayton, D. H. Epigenetics and the Evolution of Darwin's Finches. *Genome Biology and Evolution* **6,** 1972–1989. ISSN: 1759-6653 (Aug. 2014).

64. Abzhanov, A., Kuo, W. P., Hartmann, C., Grant, B. R., Grant, P. R. & Tabin, C. J. The Calmodulin Pathway and Evolution of Elongated Beak Morphology in Darwin's Finches. *Nature* **442,** 563–567. ISSN: 1476-4687 (Aug. 2006).

65. Kvon, E. Z., Kamneva, O. K., Melo, U. S., Barozzi, I., Osterwalder, M., Mannion, B. J., Tissières, V., Pickle, C. S., Plajzer-Frick, I. & Lee, E. A. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167,** 633–642. e11 (2016).

66. Cardoso-Moreira, M., Sarropoulos, I., Velten, B., Mort, M., Cooper, D. N., Huber, W. & Kaessmann, H. Developmental Gene Expression Differences between Humans and Mammalian Models. *Cell Reports* **33,** 108308. ISSN: 2211-1247 (Oct. 2020).

67. Sarropoulos, I., Sepp, M., Frömel, R., Leiss, K., *et al.* Developmental and Evolutionary Dynamics of Cis-Regulatory Elements in Mouse Cerebellar Cells. *Science* **373,** eabg4696 (Aug. 2021).

68. Krishnaswami, S. R., Grindberg, R. V., Novotny, M., Venepally, P., *et al.* Using Single Nuclei for RNA-seq to Capture the Transcriptome of Postmortem Neurons. *Nature Protocols* **11,** 499–524. ISSN: 1750-2799 (Mar. 2016).

69. Svensson, V. Droplet scRNA-seq Is Not Zero-Inflated. *Nature Biotechnology* **38,** 147–150. ISSN: 1546-1696 (Feb. 2020).

70. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nature Biotechnology* **36,** 411–420. ISSN: 1546-1696 (May 2018).

71. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* Sept. 2020. arXiv: `1802.03426 [cs, stat]`.

72. Melville, J. *Uwot* Sept. 2022.

73. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. ISSN: 1742-5468 (Oct. 2008).

74. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biology* **19,** 15. ISSN: 1474-760X (Feb. 2018).

75. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch Effects in Single-Cell RNA-sequencing Data Are Corrected by Matching Mutual Nearest Neighbors. *Nature Biotechnology* **36,** 421–427. ISSN: 1546-1696 (May 2018).

76. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r. & Raychaudhuri, S. Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nature Methods* **16,** 1289–1296. ISSN: 1548-7105 (Dec. 2019).

77. Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E. Z. & Welch, J. D. Jointly Defining Cell Types from Multiple Single-Cell Datasets Using LIGER. *Nature Protocols* **15,** 3632–3662. ISSN: 1750-2799 (Nov. 2020).

78. Wizeman, J. W., Guo, Q., Wilion, E. M. & Li, J. Y. Specification of Diverse Cell Types during Early Neurogenesis of the Mouse Cerebellum. *eLife* **8** (eds Cepko, C. L. & Bronner, M. E.) e42388. ISSN: 2050-084X (Feb. 2019).

79. Allen Institute for Brain Science (2004). Allen Mouse Brain Atlas [Dataset]. *Allen Institute for Brain Science* (2011).

80. Allen Institute for Brain Science (2004). Allen Developing Mouse Brain Atlas [Dataset]. *Allen Institute for Brain Science* (2011).

81. Visel, A., Thaller, C. & Eichele, G. GenePaint.Org: An Atlas of Gene Expression Patterns in the Mouse Embryo. *Nucleic Acids Research* **32,** D552–D556. ISSN: 0305-1048 (Jan. 2004).

82. La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., *et al.* Molecular Architecture of the Developing Mouse Brain. *Nature* **596,** 92–96. ISSN: 1476-4687 (Aug. 2021).

83. Saunders, N. R., Adam, E., Reader, M. & Møllgård, K. Monodelphis Domestica (Grey Short-Tailed Opossum): An Accessible Model for Studies of Early Neocortical Development. *Anatomy and Embryology* **180,** 227–236. ISSN: 1432-0568 (Aug. 1989).

84. Ovchinnikova, S. & Anders, S. Exploring Dimension-Reduced Embeddings with Sleepwalk. *Genome Research* **30,** 749–756. ISSN: 1088-9051, 1549-5469 (May 2020).

85. Millen, K. J., Steshina, E. Y., Iskusnykh, I. Y. & Chizhikov, V. V. Transformation of the Cerebellum into More Ventral Brainstem Fates Causes Cerebellar Agenesis in the Absence of Ptf1a Function. *Proceedings of the National Academy of Sciences* **111,** E1777–E1786 (Apr. 2014).

86. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., *et al.* SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nature Methods* **14,** 1083–1086. ISSN: 1548-7105 (Nov. 2017).

87. Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., *et al.* A Scalable SCENIC Workflow for Single-Cell Gene Regulatory Network Analysis. *Nature Protocols* **15,** 2247–2276. ISSN: 1750-2799 (July 2020).

88. Carter, R. A., Bihannic, L., Rosencrance, C., Hadley, J. L., Tong, Y., Phoenix, T. N., Natarajan, S., Easton, J., Northcott, P. A. & Gawad, C. A Single-Cell Transcriptional Atlas of the Developing Murine Cerebellum. *Current Biology* **28,** 2910–2920.e2. ISSN: 0960-9822 (Sept. 2018).

89. Schüller, U., Kho, A. T., Zhao, Q., Ma, Q. & Rowitch, D. H. Cerebellar 'Transcriptome' Reveals Cell-Type and Stage-Specific Expression during Postnatal Development and Tumorigenesis. *Molecular and Cellular Neuroscience* **33,** 247–259. ISSN: 1044-7431 (Nov. 2006).

90. Peng, J., Sheng, A.-l., Xiao, Q., Shen, L., Ju, X.-C., Zhang, M., He, S.-T., Wu, C. & Luo, Z.-G. Single-Cell Transcriptomes Reveal Molecular Specializations of Neuronal Cell Types in the Developing Cerebellum. *Journal of Molecular Cell Biology* **11,** 636–648. ISSN: 1759-4685 (Aug. 2019).

91. Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., *et al.* The Origin and Evolution of Cell Types. *Nature Reviews Genetics* **17,** 744–757. ISSN: 1471-0064 (Dec. 2016).

92. Parmigiani, E., Leto, K., Rolando, C., Figueres-Oñate, M., López-Mascaraque, L., Buffo, A. & Rossi, F. Heterogeneity and Bipotency of Astroglial-Like Cerebellar Progenitors along the Interneuron and Glial Lineages. *Journal of Neuroscience* **35,** 7388–7402. ISSN: 0270-6474, 1529-2401 (May 2015).

93. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., *et al.* The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. *Nature* **581,** 434–443. ISSN: 1476-4687 (May 2020).

94. Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., *et al.* The Human Gene Mutation Database (HGMD®): Optimizing Its Use in a Clinical Diagnostic or Research Setting. *Human Genetics* **139,** 1197–1207. ISSN: 1432-1203 (Oct. 2020).

95. Aldinger, K. A., Thomson, Z., Phelps, I. G., Haldipur, P., *et al.* Spatial and Cell Type Transcriptional Landscape of Human Cerebellar Development. *Nature Neuroscience* **24,** 1163–1175. ISSN: 1546-1726 (Aug. 2021).

96. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., *et al.* RNA Velocity of Single Cells. *Nature* **560,** 494–498. ISSN: 1476-4687 (Aug. 2018).

97. Bakken, T. E., Jorstad, N. L., Hu, Q., Lake, B. B., *et al.* Comparative Cellular Analysis of Motor Cortex in Human, Marmoset and Mouse. *Nature* **598,** 111–119. ISSN: 1476-4687 (Oct. 2021).

98. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., *et al.* Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification. *Bioinformatics* **21,** 650–659. ISSN: 1367-4803 (Mar. 2005).

99. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., *et al.* Tissue-Based Map of the Human Proteome. *Science* **347,** 1260419 (Jan. 2015).

100. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., *et al.* The Evolution of Gene Expression Levels in Mammalian Organs. *Nature* **478,** 343–348. ISSN: 1476-4687 (Oct. 2011).

101. Clark, N. M., Fisher, A. P. & Sozzani, R. in *Computational Cell Biology: Methods and Protocols* (eds von Stechow, L. & Santos Delgado, A.) 139–151 (Springer, New York, NY, 2018). ISBN: 978-1-4939-8618-7.

102. Xu, X., Stoyanova, E. I., Lemiesz, A. E., Xing, J., Mash, D. C. & Heintz, N. Species and Cell-Type Properties of Classically Defined Human and Rodent Neurons and Glia. *eLife* **7** (eds Zoghbi, H. Y. & West, A. E.) e37551. ISSN: 2050-084X (Oct. 2018).

103. Vladoiu, M. C., El-Hamamy, I., Donovan, L. K., Farooq, H., *et al.* Childhood Cerebellar Tumours Mirror Conserved Fetal Transcriptional Programs. *Nature* **572,** 67–73. ISSN: 1476-4687 (Aug. 2019).

104. Hibi, M. & Shimizu, T. Development of the Cerebellum and Cerebellar Neural Circuits. *Developmental Neurobiology* **72,** 282–301. ISSN: 1932-846X (2012).

105. Hashimoto, R., Hori, K., Owa, T., Miyashita, S., *et al.* Origins of Oligodendrocytes in the Cerebellum, Whose Development Is Controlled by the Transcription Factor, Sox9. *Mechanisms of Development* **140,** 25–40. ISSN: 0925-4773 (May 2016).

106. Sudarov, A., Turnbull, R. K., Kim, E. J., Lebel-Potter, M., Guillemot, F. & Joyner, A. L. Ascl1 Genetics Reveals Insights into Cerebellum Local Circuit Assembly. *Journal of Neuroscience* **31,** 11055–11069. ISSN: 0270-6474, 1529-2401 (July 2011).

107. Karam, S. D., Burrows, R. C., Logan, C., Koblar, S., Pasquale, E. B. & Bothwell, M. Eph Receptors and Ephrins in the Developing Chick Cerebellum: Relationship to Sagittal Patterning and Granule Cell Migration. *Journal of Neuroscience* **20,** 6488–6500. ISSN: 0270-6474, 1529-2401 (Sept. 2000).

108. Larouche, M. & Hawkes, R. From Clusters to Stripes: The Developmental Origins of Adult Cerebellar Compartmentation. *The Cerebellum* **5,** 77–88. ISSN: 1473-4230 (June 2006).

109. Redies, C., Neudert, F. & Lin, J. Cadherins in Cerebellar Development: Translation of Embryonic Patterning into Mature Functional Compartmentalization. *The Cerebellum* **10,** 393–408. ISSN: 1473-4230 (Sept. 2011).

110. Kruschke, J. K. Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science* **1,** 270–280. ISSN: 2515-2459 (June 2018).

111. Ross, S. E., McCord, A. E., Jung, C., Atan, D., *et al.* Bhlhb5 and Prdm8 Form a Repressor Complex Involved in Neuronal Circuit Assembly. *Neuron* **73,** 292–303. ISSN: 0896-6273 (Jan. 2012).

112. Britanova, O., de Juan Romero, C., Cheung, A., Kwan, K. Y., *et al.* Satb2 Is a Postmitotic Determinant for Upper-Layer Neuron Specification in the Neocortex. *Neuron* **57,** 378–392. ISSN: 0896-6273 (Feb. 2008).

113. González-Blas, C. B., Winter, S. D., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S. & Aerts, S. *SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks* Aug. 2022.

114. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion Pseudotime Robustly Reconstructs Lineage Branching. *Nature Methods* **13,** 845–848. ISSN: 1548-7105 (Oct. 2016).

115. Helms, A. W. & Johnson, J. E. Specification of Dorsal Spinal Cord Interneurons. *Current Opinion in Neurobiology* **13,** 42–49. ISSN: 0959-4388 (Feb. 2003).

116. Butts, T., Rook, V., Varela, T., Wilson, L. & Wingate, R. J. T. in *Handbook of the Cerebellum and Cerebellar Disorders* (eds Manto, M. U., Gruol, D. L., Schmahmann, J. D., Koibuchi, N. & Sillitoe, R. V.) 99–119 (Springer International Publishing, Cham, 2022). ISBN: 978-3-030-23810-0.

117. Goffinet, A. M. The Embryonic Development of the Cerebellum in Normal and Reeler Mutant Mice. *Anatomy and Embryology* **168,** 73–86. ISSN: 1432-0568 (Oct. 1983).

118. Yuasa, S., Kawamura, K., Ono, K., Yamakuni, T. & Takahashi, Y. Development and Migration of Purkinje Cells in the Mouse Cerebellar Primordium. *Anatomy and Embryology* **184,** 195–212. ISSN: 1432-0568 (May 1991).

119. Northcott, P. A., Shih, D. J. H., Peacock, J., Garzia, L., *et al.* Subgroup-Specific Structural Variation across 1,000 Medulloblastoma Genomes. *Nature* **488,** 49–56. ISSN: 1476-4687 (Aug. 2012).

120. Roussel, M. F. & Stripay, J. L. Modeling Pediatric Medulloblastoma. *Brain Pathology* **30,** 703–712. ISSN: 1750-3639 (2020).

121. Rio, M., Royer, G., Gobin, S., de Blois, M., *et al.* Monozygotic Twins Discordant for Submicroscopic Chromosomal Anomalies in 2p25.3 Region Detected by Array CGH. *Clinical Genetics* **84,** 31–36. ISSN: 1399-0004 (2013).

122. Oikonomakis, V., Kosma, K., Mitrakos, A., Sofocleous, C., *et al.* Recurrent Copy Number Variations as Risk Factors for Autism Spectrum Disorders: Analysis of the Clinical Implications. *Clinical Genetics* **89,** 708–718. ISSN: 1399-0004 (2016).

123. Reim, K., Wegmeyer, H., Brandstätter, J. H., Xue, M., Rosenmund, C., Dresbach, T., Hofmann, K. & Brose, N. Structurally and Functionally Unique Complexins at Retinal Ribbon Synapses. *Journal of Cell Biology* **169,** 669–680. ISSN: 0021-9525 (May 2005).

124. Sousa, A. M. M., Zhu, Y., Raghanti, M. A., Kitchen, R. R., *et al.* Molecular and Cellular Reorganization of Neural Circuits in the Human Lineage. *Science* **358,** 1027–1032 (Nov. 2017).

125. Wang, Z.-Y., Leushkin, E., Liechti, A., Ovchinnikova, S., *et al.* Transcriptome and Translatome Co-Evolution in Mammals. *Nature* **588,** 642–647. ISSN: 1476-4687 (Dec. 2020).

126. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal* **8,** 289–317. ISSN: 2073-4859 (Aug. 2016).

127. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8,** 281–291.e9. ISSN: 2405-4712 (Apr. 2019).

128. Melville, J., Lun, A. & Djekidel, M. N. Uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. *R package version* **15** (2020).

129. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. ISSN: 1742-5468 (Oct. 2008).

130. Young, M. D. & Behjati, S. SoupX Removes Ambient RNA Contamination from Droplet-Based Single-Cell RNA Sequencing Data. *GigaScience* **9,** giaa151. ISSN: 2047-217X (Nov. 2020).

131. Giorgino, T. Computing and Visualizing Dynamic Time Warping Alignments in R: The Dtw Package. *Journal of Statistical Software* **31,** 1–24. ISSN: 1548-7660 (Aug. 2009).

132. Baglama, J. in *Handbook of Big Data* (Chapman and Hall/CRC, 2016). ISBN: 978-0-429-16298-5.

133. Kumar, L. & E. Futschik, M. Mfuzz: A Software Package for Soft Clustering of Microarray Data. *Bioinformation* **2,** 5–7. ISSN: 0973-2063 (May 2007).

134. Brunson, J. Ggalluvial: Layered Grammar for Alluvial Plots. *Journal of Open Source Software* **5,** 2017. ISSN: 2475-9066 (May 2020).

135. Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D. & Cooper, D. N. The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies. *Human Genetics* **136,** 665–677. ISSN: 1432-1203 (June 2017).

136. Gröbner, S. N., Worst, B. C., Weischenfeldt, J., Buchhalter, I., *et al.* The Landscape of Genomic Alterations across Childhood Cancers. *Nature* **555,** 321–327. ISSN: 1476-4687 (Mar. 2018).

137. Wickham, H., Averick, M., Bryan, J., Chang, W., *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4,** 1686. ISSN: 2475-9066 (Nov. 2019).

138. Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., *et al.* Orchestrating Single-Cell Analysis with Bioconductor. *Nature Methods* **17,** 137–145. ISSN: 1548-7105 (Feb. 2020).

139. Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C. & Macosko, E. Z. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177,** 1873–1887.e17. ISSN: 0092-8674 (June 2019).

140. Wickham, H. Elegant Graphics for Data Analysis (Ggplot2). *Applied Spatial Data Analysis R* (2009).

141. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing High-Throughput Sequencing Data in Python with HTSeq 2.0. *Bioinformatics* **38,** 2943–2945. ISSN: 1367-4803 (May 2022).