Miguel Monteiro Pacheco **Characterization of the choroid plexus cells transcriptome during development**

UMinho | 2021

Universidade do Minho
Escola de Engenharia

Miguel Monteiro Pacheco

Characterization of the choroid plexus cells transcriptome during development

dezembro de 2021

Universidade do Minho
Escola de Engenharia

Miguel Monteiro Pacheco

Characterization of the choroid plexus cells transcriptome during development

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação da
Doutora Ana Falcão
e do
Professor Doutor Miguel Rocha

dezembro de 2021

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Licença concedida aos utilizadores deste trabalho

Agradecimentos

Este trabalho foi apoiado por uma bolsa da Fundação “la Caixa” (ID100010434) com o código de bolsa LCF / BQ / PI19 / 11690005, e pela Swedish Research Council (grant no. 2019-02030).

Agradecemos às instalações de Eukaryotic Single Cell Genomics (ESCG) em Estocolmo financiado pela SciLifeLab, KICore e StratRegen, onde os dados de RNA-seq de célula única foram gerados, e ao Professor João Sousa e à sua equipa Diana Afonso, Ana Veloso e Fernanda Marques do ICVS, Universidade do Minho, que gerou os dados de Bulk RNA-seq.

Em seguida gostaria também de agradecer ao pessoal do laboratório de Biosystems/ OmniumAI por me acolherem na fase final e me darem um bom ambiente de trabalho que foi crucial para acabar. Em particular, ao meu antigo colega Fernando Cruz que sempre se demonstrou disponível para me ajudar.

Um agradecimento especial também para a Rita por me ter ouvido e ter estado sempre lá para mim durante este período de tese, lidar bem com o meu pânico e dar-me a força necessária para continuar sem desistir. Ao mesmo tempo, ela e o grupo ajudaram-me também a abstrair-me quando era necessário e a proporcionar sempre bons momentos. Neste mesmo grupo queria também agradecer a todos os que me mantiveram sempre atento e a puxar pela cabeça no xadrez, Braga e Paulo, senti que conseguia manter muito mais o foco e a concentração por mais tempo por causa disto.

Um especial obrigado também ao meu grande amigo “Gajo” que infelizmente/ felizmente partiu para o Algarve, espero que a tua carreira tenha sucesso aí, e que te tentes divertir sem mim, jogar xadrez sem tabuleiro físico vai ser sem dúvida alguma, menos divertido.

Acho que nunca vou ter obrigados suficientes para a Ana e muito menos para a Lili (e sei que disse muitos). Foram sem dúvida imprescindíveis durantes todos estes meses, e tenho a certeza absoluta de que nada disto teria sido possível se não fossem vocês, a criarem um ótimo ambiente e manterem-me interessado neste projeto que apesar das dificuldades foi sem dúvida uma experiência incrivelmente positiva, e definitivamente marcante na minha vida. Nunca vos esquecerei e espero não perder o contacto.

Por último, mas não por menos, gostava de agradecer aos meus pais e à Renata, vocês mais do que ninguém lidaram com a minha pressão e tiveram a paciência e compreensão para não desistirem de mim... Sei que libertei bastante pressão em vocês e espero poder compensar-vos por isso, tenho a certeza que eu nestes últimos tempos, não me aturava, mas vocês efetivamente fizeram-no, muito obrigado.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Abstract

The choroid Plexus (CP) is a brain tissue responsible for the production and secretion of the cerebrospinal fluid (CSF). It lays at the interface of the peripheral blood and the brain, forming the blood-CSF barrier, and displays a connected monolayer of epithelial cells that selects the contents from the blood that reach the brain playing a key role in brain homeostasis. In order to have a deeper understanding about the role of the CP in brain development, it is necessary to investigate the CP cell type composition and cellular states throughout several developmental timepoints, which can be performed by transcriptomic analysis.

Single-cell RNA-sequencing (scRNA-seq) is a revolutionary technology for transcriptome analysis as it allows a high throughput single-cell gene expression profiling from a tissue. This technique also provides data to infer cellular differentiation and future transcriptomic states.

In this work, the CP transcriptome was analysed in three different timepoints (two postnatal stages and one adult) by two techniques bulk RNA-seq and scRNA-seq.

While scRNA-seq analysis in the CP allowed the identification of all cell types that compose the CP as well as some differences in the expression profile between the different CP stages, bulk RNA-seq data allowed an overall analysis of the tissue transcriptomics exhibiting a more pronounced differential gene expression analysis between CP stages.

Bulk RNA-seq data analysis demonstrated that CP cells at earlier stages are enriched in genes associated with cell division (*Tuba1a*, *Cul*) and cell adhesion (*Tubb*, *Actb, Col4a1*) while in adulthood CP was enriched in genes associated to lipid and mitochondrial pathways, such as Ascl3 and Cox8b, respectively.

Importantly, scRNA-seq data analysis not only confirmed part of the bulk RNA-seq data but also lead to the identification of a subgroup in epithelial cells that expressed ciliogenesis genes in the early stages of development. Furthermore, the differentially expressed genes uncovered by bulk RNA-seq were assigned to cell types in scRNA-seq.

This study unravels several pathways enriched in developmental stages that will be investigated in the future for their role in brain development modulation.

Keywords

Bulk RNA-seq; Choroid Plexus; *Mus musculus*; *Rattus norvegicus*;; Single-cell RNA-seq;

Resumo

O plexo coróide (PC) é um tecido cerebral responsável pela produção e secreção do líquido cefalorraquidiano (LCR). É situado na interface do sangue periférico e do cérebro, formando a barreira sangue-LCR, e exibe uma monocamada conectada de células epiteliais que seleciona o conteúdo do sangue que chega ao cérebro, desempenhando um papel fundamental na homeostase cerebral. De modo a ter um entendimento mais profundo sobre o papel do CP no desenvolvimento do cérebro, é necessário investigar a composição dos tipos de células do CP e os seus respetivos estados celulares ao longo de vários estágios de desenvolvimento, que podem ser realizados por análise transcriptómica.

O sequenciamento de RNA de célula única (scRNA-seq) é uma tecnologia revolucionária para análise de transcriptoma, pois permite um gerar um perfil da expressão génica de cada célula de um tecido, com elevado rendimento. Esta técnica também fornece dados para inferir a diferenciação celular e futuros estados transcriptómicos.

Neste trabalho, o transcriptoma do PC foi analisado em três diferentes momentos (dois estágios pós-natais e um adulto) por duas técnicas bulk RNA-seq e scRNA-seq.

Enquanto a análise de scRNA-seq no PC permitiu a identificação de todos os tipos de células que compõem o PC, bem como algumas diferenças no perfil de expressão entre os diferentes estágios do PC, os dados em Bulk RNA-seq permitiram uma análise geral da transcriptómica do tecido exibindo uma análise diferencial mais acentuada de expressão génica entre os vários estágios do PC.

A análise de dados Bulk RNA-seq demonstrou que as células PC em estágios iniciais são enriquecidas em genes associados à divisão celular (*Tuba1a*, *Cul*) e adesão celular (*Tubb*, *Actb, Col4a1*) enquanto na idade adulta a CP foi enriquecida em genes associado às vias lipídicas e mitocondriais, como *Ascl3* e *Cox8b*, respectivamente.

É importante realçar que a análise de dados de scRNA-seq não só confirmou parte dos dados de Bulk RNA-seq, mas também levou à identificação de um subgrupo em células epiteliais que expressaram genes de Ciliogénese nos estágios iniciais de desenvolvimento. Além disso, os genes diferencialmente expressos descobertos por Bulk RNA-seq foram atribuídos a tipos de células em scRNA-seq.

Este estudo pretende revelar vários caminhos enriquecidos em estágios de desenvolvimento que serão investigados no futuro por seu papel na modulação do desenvolvimento do cérebro.

Palavras-Chave

Bulk RNA-seq; *Mus musculus*; Plexo Coróide; *Rattus norvegicus*; Single-cell RNA-seq;

Contents

List of Figures

List of Tables

List of Acronyms

BBB	Blood Brain Barrier
BCmvn	Mean-Variance Normalized Bimodality Coefficient
BCSFB	Blood-cerebrospinal fluid barrier
CCA	Canonical Correlation Analysis
cDNA	Complementary DNA
CNS	Central Nervous System
CP	Choroid Plexus
CPEC	Choroid Plexus Epithelial Cells
CSF	Cerebrospinal Fluid
DEA	Differential Expression Analysis
DGE	Differential Gene Expression
DGM	Deep Grey Matter
DNA	Deoxyribonucleic Acid
ES	Enrichment Score
FACS	Fluorescent Activated Cell Sorting
FC	Fold Change
FDR	False Discovery Rate
FWER	Family Wise Error Rate
GEMs	Gel beads in Emotion
GLM	Generalized Linear Model
GSEA	Gene Set Enrichment Analysis
HTS	High Throughtput Senquencing
mRNA	Messenger RNA
MS	Multiple Sclerosis
NES	Normalized Enrichment Score
NPC	Neural Progenitor Cells
NSC	Neural Stem Cells

OPC	Oligodendrocyte Percursor Cell
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PCs	Principal Components
PRL	Prolactin hormone
QC	Quality Control
RNA	Ribonucleic acid
rRNA	Ribossomal RNA
scRNA-seq	Single-Cell RNA-sequencing
SNN	Shared Nearest Neighbor
STAR	Spliced Transcripts Alignment to a Reference
SVZ	Sub-Ventricular Zone
tRNA	Transfer RNA
T-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifier
VST	Variance Stabilizing Transformation

1. Introduction

1.1 Choroid Plexus (CP)

The CP is a highly vascularized brain tissue that is located in the brain ventricles [1]. There are four ventricles in the brain in which the CP are founded, two lateral that join on the third ventricle which is more central and is connected to the fourth ventricle through the cerebral aqueduct of the brain (Fig. 1A). The CPs across ventricles (Fig. 1B-D) have very similar structure, they are constituted by a single layer of epithelial cells joined by tight junctions forming the blood-cerebrospinal fluid barrier (BCSFB) [2].

The brain borders, the blood brain barrier (BBB) and BSCFB, restrict most of the immune cells migration from the periphery to the central nervous system (CNS) allowing to preserve CNS homeostasis and correct function of brain cells [3].

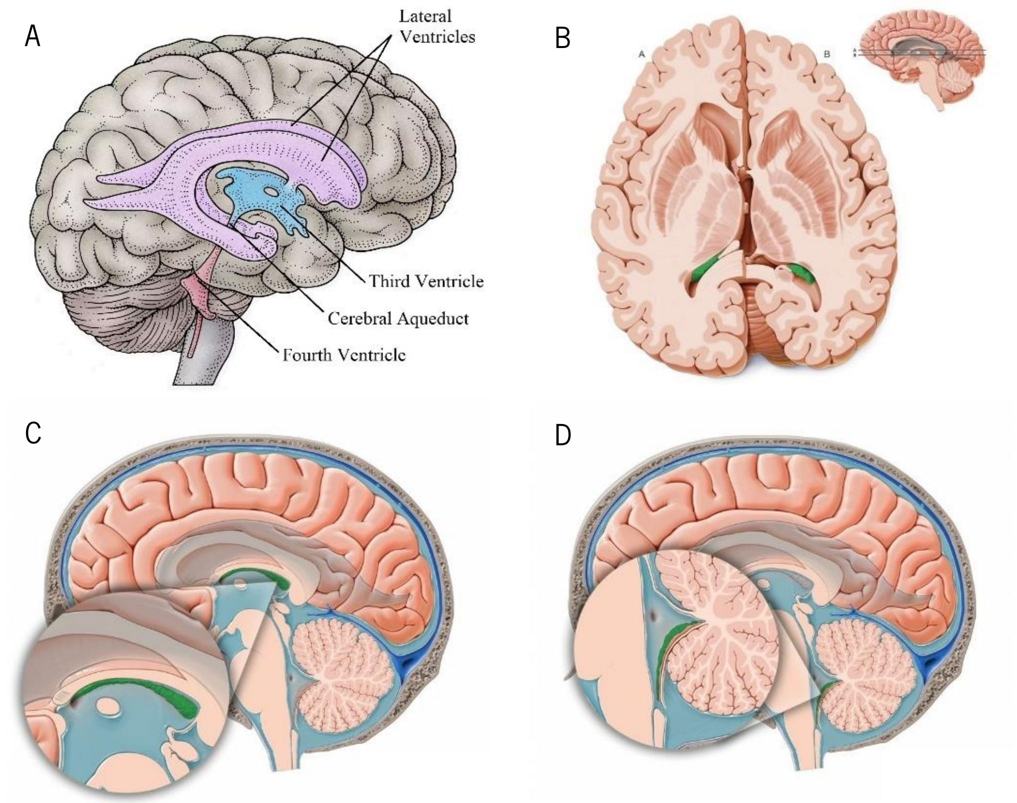

Figure 1 | *Schematic representation of the CP in the human brain. A – Illustration of the four ventricles location in the brain; B – Axial plane of the brain and site of the CP location (green) in the lateral ventricles; C – Sagittal plane of the brain and location of the CP (green) in the third ventricle; D – Sagittal plane of the brain and location of the CP in the fourth ventricle* [4][5].

The CP epithelium has associated epiplexus cells (macrophage-like cells) that are in direct contact with the CSF, whereas the CP basolateral membrane is laid over a stromal core comprised of different cell types, including fibroblasts, immune cells and numerous fenestrated capillaries (Fig.1A) [6]. The CP epithelial cells hold one of the most important functions to maintain brain homeostasis, they produce most of the CSF [6][7][8]. The CSF carries mostly inorganic ions, lipids and glucose, identical to blood plasma. Furthermore, it has a minor amount of proteins [9] and immune cells originated in the CP [10]. These components have small fluctuations, which suggests several CSF regulation mechanisms [6]. On other hand, epithelial cells also mediate the transport of metabolic waste and other molecules out of the CNS.

The CSF production relies on the unique polarity of the choroid plexus epithelial cells (CPEC), this process is proved to have an high dependency on HCO_3^-[6][11], as well as, several ion transporters in the CPEC membranes such as, the Na^+-K^+-ATPase and NKCC1 since Na^+ is quantitatively the most important ion transported [11]. CPEC also display an high water permeability due to the aquaporin-1 (AQP1) being highly expressed (Fig. 2B) [12].

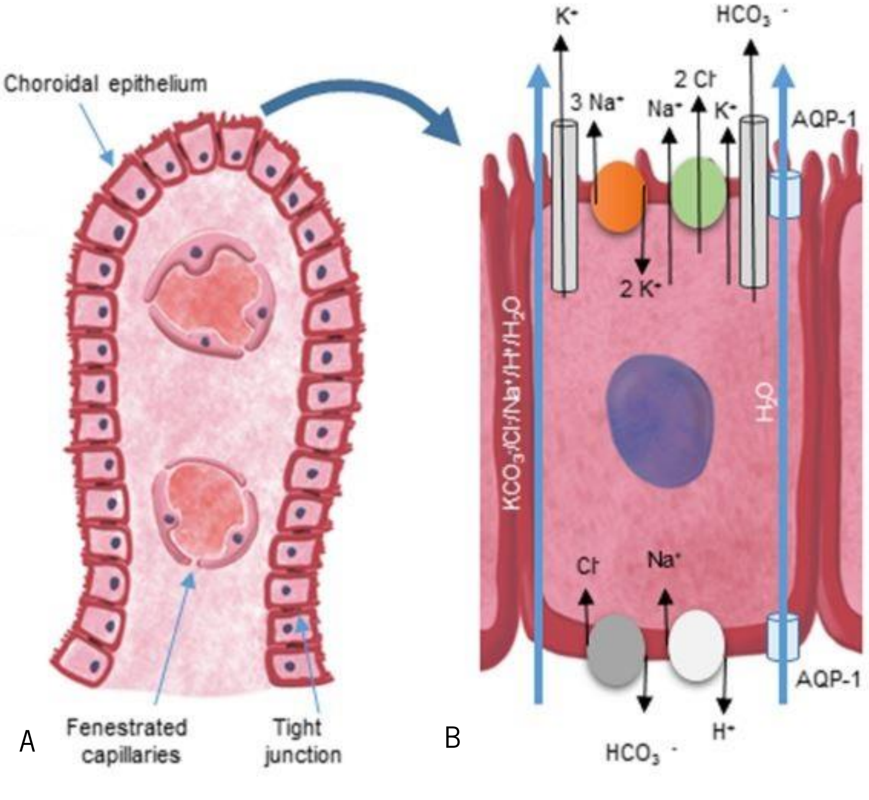

Figure 2 | *Schematic representation of CP cells. A – Scheme of the CP choroidal epithelium covered by a single layer of cuboidal epithelial cells and fenestrated capillaries in the stromal core. ; B – Explanation of the ion trades that occur in the CP epithelium cells between Na^+, K^+, HCO_3 and Cl; Presence of AQP-1 in the membrane also allows the basolateral membrane to regulate the water flow across the choroid epithelia* [13].

1.1.1 The development of the CP

After the neural tube formation, the embryonic precursor of the CNS, the CP stroma derives from invaginations of mesenchymal stem cells, that help the formation of cerebral ventricles, and the CPEC are derived from neuropithelium with roof plate origins [8][14]. Tipically, the first CP appears in the fourth ventricle, followed by the lateral and lastly the third ventricle [15]. The time at which it appears varies between species, depending on the time of gestation and brain growth rate [15]. Genetic experiments in the mouse embryo reveal a central role for the secreted morphogen Sonic Hedgehog (Shh), which belongs to a signaling pathway involved in embryonic cells differentiation, in coordenating the CPEC development alongside with vasculature [16]. In fact, the absence of Shh causes an underdeveloped structure deficient in CPEC and vasculature evidencing its importance in the co-development of two separate cell lineages in CP morphology [8][16].

Despite the importance of the CSF production by the CP for the brain homeostasis, other functions of the CP have been overshadow by this feature. Through the last decades, breakthrough findings are significantly changing the perspective on the role of the CP in CNS dynamics both in development and disease.

1.1.2 The CP and its role in brain development

CP is able to modulate the fate of neural stem cells (NSC) from the subventricular zone (SVZ). Through an experiment where stem cells were bathed with embryonic CSF it was possible to conclude that the embryonic CSF contributes to the development and growth of neural stem cells [17]. CSF-derived protein signals fluctuate with age, for instance, the CSF-insulin-like growth factor 2 (IGF2), a known factor that stimulates neural stem cells division, exhibits higher levels during brain development [17]. Several similar findings demonstrate that the embryonic CP-CSF system actively distributes not only growth factors and morphogens but also cytokines, binding proteins, extracellular matrix proteins and many other factors, thus instructing the cerebral cortical development [8]. Lastly, secreted factors from the lateral ventricle CP proved to directly regulate the behaviour of stem cells and their respective progeny in the adult SVZ. The release of bone morphogenetic protein 5 (BMP5) and IGF-1 decreases with age in the lateral ventricle CP which can be an important factor for the lack recruitment of adult NSC in an aged brain [18]. Nevertheless, how the CP secretome can influence NSC or neural progenitor cells (NPC) progeny remains largely unknown.

1.1.3 The CP and its role in inflammatory diseases

Inflammation is the main cause of neuronal death in the CNS. Its origin can be due to, infections, trauma or neurodegenerative diseases, such as multiple sclerosis (MS). At the BCSFB barrier, T cell can cross to CSF and invade the brain [19]. A study in rats with traumatic brain injury observed neutrophils (granulocyte) infiltration via CP of lateral ventricles, due to the expression of CXC chemokines, such as, CXCL1, CXCL2 and CXCL3 in the tissue [20]. Reports also confirm the accumulation of monocytes and neutrophils in the CSF through the paracellular route which sugests the CPEC tight junctions disrupture in injuries and infections of the brain (Fig. 3) [21]. Adhesion molecules such as VCAM-1, ICAM-1, P-selectin and E-cadherin are constitutively expressed by CPEC on their surface preventing the leukocytes exit from the CP vasculature/ stroma into the ventricular lumen [22]. Nevertheless, due to increased MCP-1 production and secretion to CSF by the CP, leucocyte migration through the barrier is facilitated [21]. The presence of immune cells in CP is not necessarily bad, in fact, most of the T-cells were found to be in the CP, CSF and meningeal membranes which cover the brain [23]. Furthermore, the presence of these cells in the CP proved to regulate immune cell trafficking through the production of interferon-γ (IFN-γ) which upregulates the ICAM-1 and VCAM-1 [7].

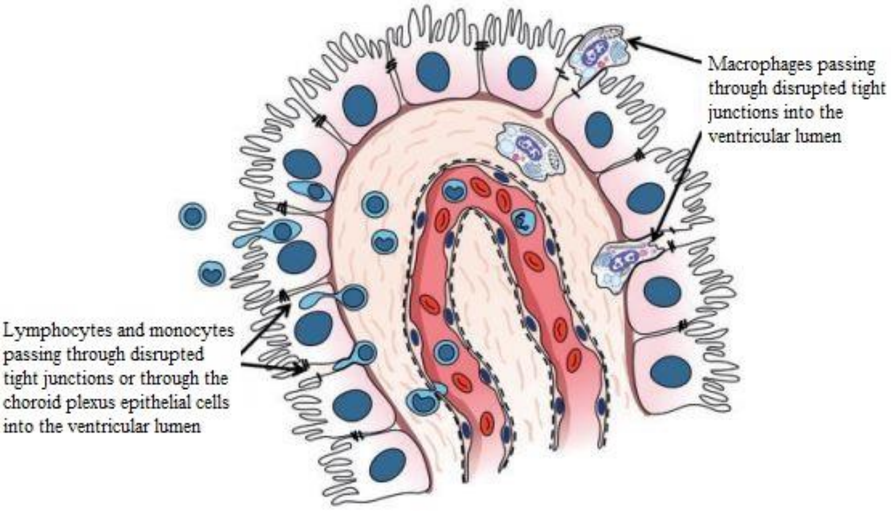
Macrophages passing through disrupted tight junctions into the ventricular lumen
Lymphocytes and monocytes passing through disrupted tight junctions or through the choroid plexus epithelial cells into the ventricular lumen

Figure 3 | *Illustration on the infiltration of stromal macrophages, lymphocytes, and monocytes into the ventricular lumen through the CP epithelial cells disrupted tight junctions/ compromised brain-cerebrospinal-fluid-blood barrier* [26].

Other immune cells harbored in the CP are also associated with the upregulation of factors such as brain-derived neurotrophic factor (BDNF) and insuline-like growth factor (IGF-1), that benefit the neural parenchyma in disease conditions [24]. However, these cells appear to remain inactive in the CP sugesting a brain inherent ability to control their autoimmune activity [7]. For instance, some studies observed that certain cytokines such as CCL11 can impair adult neurogenesis and neurotrophic factors, and possibly induce CNS microglia alteration into a proinflamattory state [25], which can indirectly increase inflammation in autoimmune diseases. As mentioned earlier, the CPEC have associated epiplexus cells (macrophage like cells) that support the microvilli structure (Fig. 4). Epiplexus cells express not only complement type 3 receptors (CR3), like the other macrophages, but they also express MHC-I and MHC-II molecules and can serve as antigen-presenting cells for lymphocytes [26].

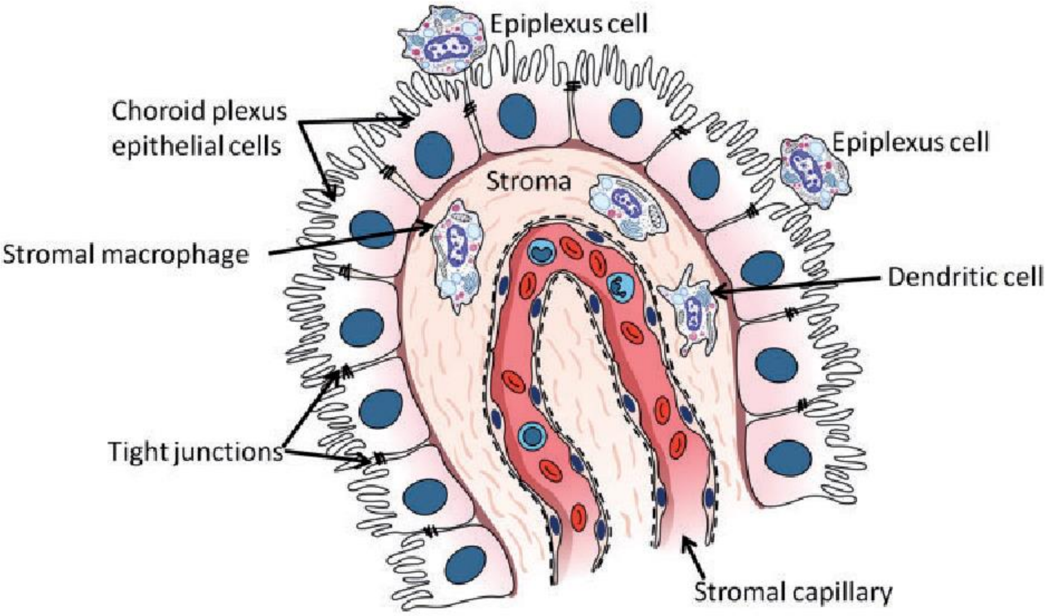

Figure 4 | *Illustration of the CP in the lateral ventricle showing the CPEC and other associated features/ cells, such as the epiplexus cells* [26].

In MS, it is hypothesised that the CP is a gateway for immune cells infiltration onto the CNS through the CSF [22]. This is supported by the damaged deep grey matter (DGM) near the ventricles in the early stages of the disease since it can be one of the primary spots for the invasion of lymphocyte from the CSF [27].

In sum, the CP is a key structure for early brain development, maintenance of brain homeostasis, and a primary target for disease prevention and treatment. However, little is known regarding the molecular mechanisms governing these functions of the CP and its cellular networks.

1.2 Bulk RNA-seq

For decades, several genes have been studied individually through techniques such as *in situ* hybridization and RT-PCR allowing a spatial-temporal perspective on their expression levels and patterns. However, the advents of micro-array technology and RNA sequencing (RNA-seq) allowed the increase of studies at a genome-wide scale of a determined tissue or organ [28]. Through techniques like bulk RNA-seq, it is now possible to collect an unbiased transcriptome, the sum of all RNA transcripts, from a determined sample.

The analysis of gene expression by molecular biologists allows to dictate what cells are doing or can react to [29]. The first RNA-seq protocols allowed the sequencing of complementary DNA (cDNA) on a large scale from a determined cell population. Nowadays, the system has been optimized, not only through better quality materials, but also through different types of materials, and further matured protocols. The basic procedures for a standard RNA-seq experiment (Fig. 5) start from the extraction and purification of RNA from a sample, followed by its enrichment. RNA enrichment mostly consists in the Poly(A) capture, that allows the selection of polyadenilated RNAs, typically messenger RNA (mRNA), as well as the depletion of the ribossomal and transfer RNAs (rRNA and tRNA) that constitute the majority of the sample (aproximately 95%) [30]. RNAs of interest are then chemically or enzimatically fragmented into a smaler size to be sequenced. Since current systems for sequencing only sequence DNA, the targeted RNAs are converted into cDNA and ligated with adapter sequences in either the 3’ and 5’ ends of the double-stranded cDNA. Before sequencing, these fragments are amplified via polymerase chain reaction (PCR) using the parts of the adapter sequence as primers. Improvements in this technique were made in the length of reads allowed, due to better sequencing machines, although the number of reads still ranges from 10 to 100 million for experiment, but lately, with a trend to deep sequencing, thus reducing the amount of errors from the process and fundamently detect rare clonal types, cells, among others.

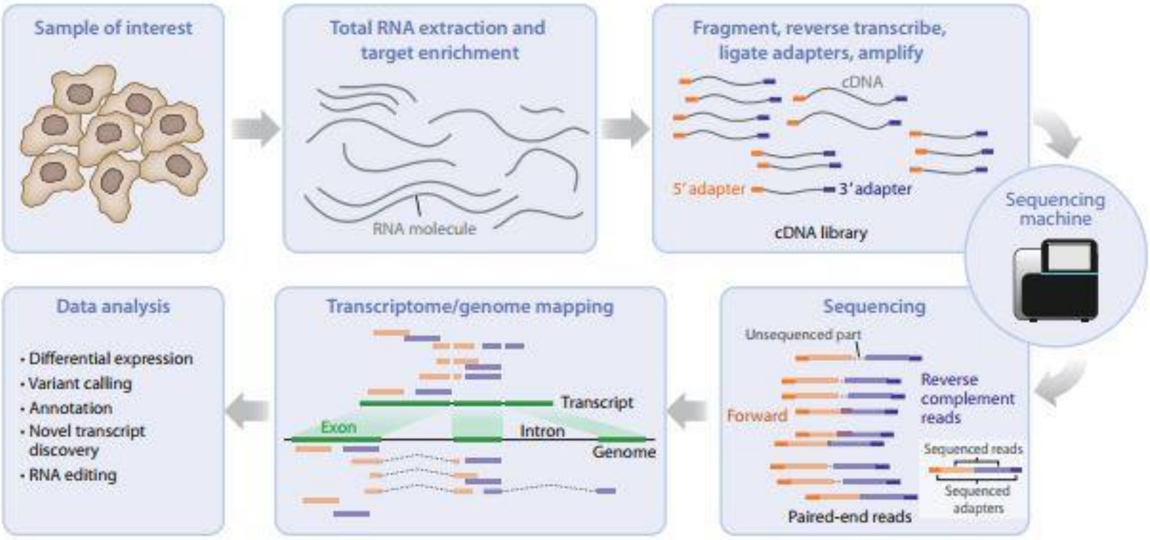

Figure 5 | *General overview of a RNAseq protocol. After the RNA extraction and amplification, the fragments are sequenced and mapped generating data capable of being analysed. Data analysis depends on the experiment goal* [29].

The RNA-seq popularity relies on its large number of aplications, such as genome annotation. Transcriptomes of model organisms and humans are not yet complete and the continued use of this method only increases the quality of these organisms libraries, allowing more cohesive information. Another aplication also includes the comparison of genes/ transcripts expression between different tissues, cell types, as well as stimulation conditions, disease states and growth conditions, allowing the identification of genes that change in expression to understand molecular pathways or, for instance, disease stages [29]. Overall, RNA-seq has a simple workflow, clear designed pipelines and projects, manageable outputs and several bioinformatics tools that facilitate the analysis and comprehension of its data, mainly due to the constant optimization of the method. However, it does produce very big output files that require a high volume of storage, and further powerful tools to analyse the data. Repetitive sequences, high sequence similarity between alternative spliced isoforms and non availability of genome references of certain organisms also cause difficulties in the viability of the results it produces. Lastly, other factors such as price and data processing time will tend to decrease over the time, increasing the praticability of RNA-seq [31].

1.3 Single cell RNA-seq

Despite the many uses for bulk RNA-seq, certain applications require single cell resolution, especially when studying samples provenient from heterogenous tissues or consist of more than one cell type. Even though bulk RNA-seq can computationally estimate the cell composition of a determined sample, scRNAseq offers a more accurate and reliable information, allowing new cell discovery or even the performance of a cell-type transcriptome analysis [32].

Cells are the smalest functional units in the body, thus it is imperative to develop techniques that allow the study of all of its aspects. Due to the advance in technologies of barcoding and decrease of intense labor required by micromanipulation, such as the fluorescent activated cell sorting (FACS) technique, it is possible to deposit single cells into micro-wells [33]. This improvement, together with new molecular strategies to individually barcode and amplify the transcriptome of each cell and further advances in microfluidics, allowed the automation of single cell docking and their respective molecules reducing the consumption of reagents and increasing their sensivity [34]. ScRNA-seq arises with an increased cellular throughput turning it into a desirable and powerful system to analyse heterogenous tissue [33], despite its increased cost. ScRNA-seq methodologies are dependent on the biological question and biological material.

Nevertheless, the most widespread method is the droplet-based scRNA-seq (Fig. 6) which will be further discussed below. For all techniques of scRNA-seq, it is imperative to start with a good sample preparation. This step can take several months to be optimized according to each cell type. Generally, the sample is dissociated into a single cell suspension (using mechanic and/or enzymatic dissociation), and then purified to remove dead cells or enrich for the cells of interest, via FACS, for example. After purification cells are individually encapsuled into gel beads and emerged in partitioning oil forming gel beads in emulsion (GEMs) (Fig. 5.2 – 10x Chromium). As the third step, each cell is barcoded, as well as each of its molecules, identified by the unique molecular identifier (UMI), followed by later amplification for sequencing. After the amplification, either by PCR or RT-PCR, the molecules are sequenced and, lastly, processed by bioinformatics analytical tools that have been developed specifically to manage and interpret scRNA-seq data [35].

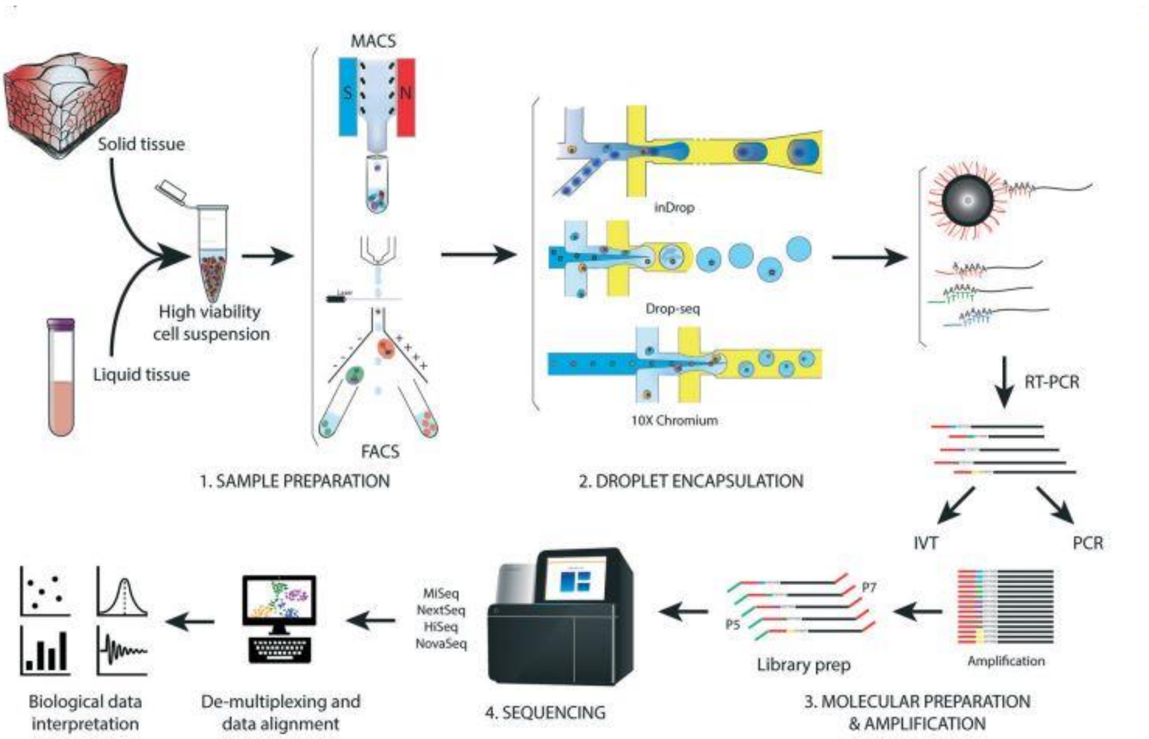

Figure 6 | *General overview of a scRNAseq protocol. After sample preparation the cells are encapsulated in droplets and its content barcoded and amplified. Molecules are then sequenced and analysed by bioinformatic tools, ready to be interpreted* [33].

The aplications of scRNA-seq can distinctively identify the cell composition of tissues, including small and unknown populations [36], as well as to allow the better understanding of cell differentiation, activation and polarisation [37][38]. Furthermore, it can also help to understand cell-cell comunication through receptor-ligand networks in healthy cells and their response to genetic manipulation or drugs [39][40]. Despite the numerous advantages and powerfull information given by the scRNA-seq system, it still includes limitations such as, the difficulty to distinguish between technical noise and biological variability for low transcripts, and to maintain strand specificity and detect isoforms in parallel, furthermore, post-transcriptional RNA modifications and other RNA editing events are also not explored in scRNAseq [41]. However, several bioinformatics tools can help and minimize these issues through the creation of quality control (QC) points along the data analysis, as well as, the creation of models that can simulate these post-transcriptional events [42]. The continuous innovation of scRNA-seq technologies and its respective bioinformatics approaches can promote biological and clinical research and provide new insights into heterogenous tissue and cell dynamics [43].

1.4 Motivation and Objectives

The CP has a central, yet poorly investigated, role in the modulation of brain cell dynamics such as the formation of new cell types from progenitor cells lying at the SVZ. It is known that CP regulates both embryonic and adult neural stem and precursor cells through different mechanisms. Determining the CP transcriptome profile in different developmental stages will help to infer about CSF composition and its impact on the processes that lead to brain cell fate decisions during development. Furthermore, in order to successfully predict the behaviour and analyse the CP response in diseases progression it is necessary to first describe its development to better understand CP physiological functionality.

The scRNA-seq technology can greatly improve this understanding, specially when used alongside the Bulk RNA-seq, providing not only a general overview of the tissue but also a more detailed analysis about the multiple cell types present and their specific development.

As such, the main objective of this thesis is to characterize the transcriptomic profile of CP cells in different developmental time points using two methodologies: single-cell RNA seq and bulk RNA-seq.

2. Materials and Methods

2.1 Bulk RNA-seq analysis

Bulk RNA-seq was performed in healthy rats at five postnatal time points: P1 (postnatal day 1), P4, P7, P10 and P60, with an n=3 for each time point, with the exception of P60 which has n=2. The chosen time points hold important hallmarks for brain development such as, NSC differentiation, oligodendrocyte precursor cell (OPC) formation and neurogenesis. Bulk RNA-seq sample information is presented in table 1, where each sample has a unique code and barcode, as well as the concentration of RNA. 28S:18S is the ratio of 28S and 18S rRNA indicating the quality of the RNA where the ratio 2:1 shows non-degraded RNA. It also displays the RQI, the RNA quality index (it scales from 1 to 10, being 10 a highly intact RNA and 1 highly degraded RNA).

Table 1 | *Bulk-RNAseq samples information.*

Age	Code	Sample number	Concentration (ng/μL)	28S18S	RQI	Barcode
P1	DA.15.01.059	1	140.92	1.64	9.9	BC01
	DA.15.01.061	2	156.65	1.71	9.9	BC02
	DA.15.01.062	3	56.3	1.65	9.9	BC03
P10	DA.15.01.052	10	178.68	1.47	9.8	BC037
	DA.15.01.054	11	381.39	1.45	9.6	BC038
	DA.15.01.056	12	484.11	1.41	9.7	BC039
P60	DA.15.01.033	14	617.69	1.26	9.6	BC050
	DA.15.01.035	15	457.79	1.23	9.4	BC051

Illumina sequencing was the used, as the technology to obtain the bulk data.

This unpublished data was generated by Diana Afonso, Ana Veloso, Fernanda Marques, and João Carlos Sousa from ICVS, University of Minho.

2.1.1 Bulk RNA-seq data pre-processing

Typically, pre-processing bulk RNA-seq data follows a determined protocol. The following section will identify a typical protocol for this task (Fig. 7) and describe the used protocol for the pre-processing of bulk RNA-seq sample files.

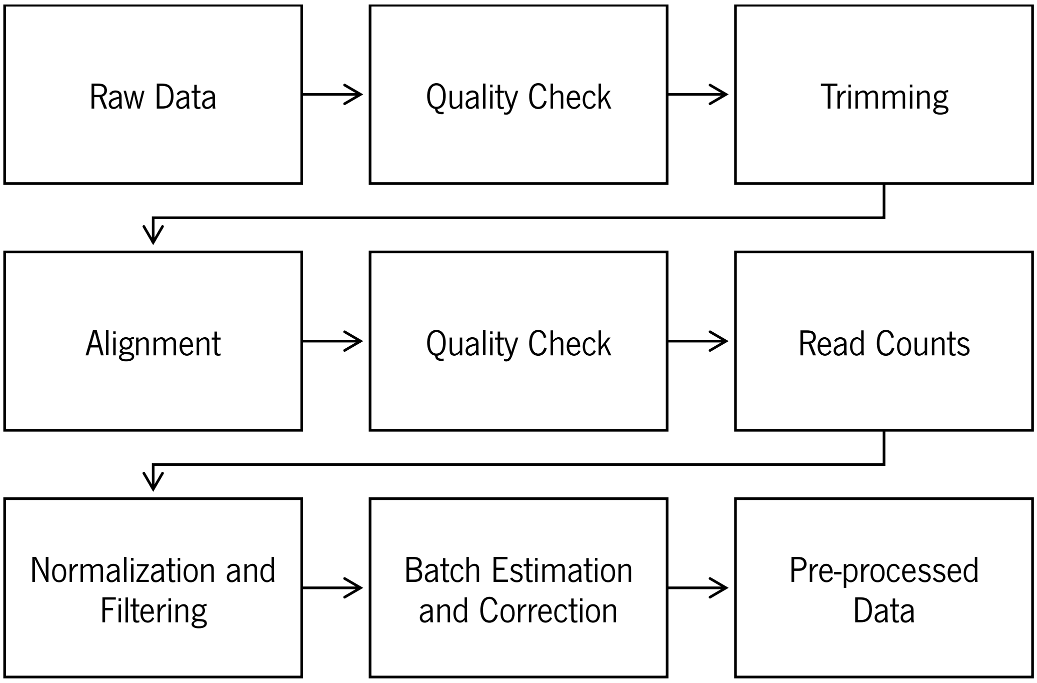

Figure 7 | *General overview of a common pre-processing for bulk RNAseq data. Several steps can have multiple bioinformatic tools associated as well as multiple options for the same effect.*

Quality control

The sample files were obtained in the BAM format, a non-human readable format, mainly used for sequence data storage since its compressed. Therefore, the first step is to convert them into readable files. For that purpose, all sample files were converted to the FASTQ (FQ) format through SAMTools (https://quay.io/repository/biocontainers/samtools, tag: 1.12–hd5e65b6_0). This output format allows the first round of quality control check to be made through FastQC (https://quay.io/repository/biocontainers/fastqc, tag: 0.11.9—0). After the first quality control check, Trim Galore (wrapper of Cutadapt and FastQC) ensured the adapter and quality trimming of the samples (https://quay.io/repository/biocontainers/trim-galore, tag 0.6.6—0).

Trimming is an essential step in pre-processing data, that allows the removal of the sequencing adapters, polyA tails, as well as reads with poor quality that can affect downstream analysis, Trim Galore used the Cutadapt version 2.10, in the single-end trimming mode and removes reads with Phread Score lower than 20 or with less than 20 bases. The Phred quality score indicates a probability of a determined base being correctly assigned. The scores range from 2 to 40 and higher scores indicate greater confidence in accuracy [44]. Furthermore, it also recognizes the Illumina adapters used for the sample sequencing allowing their removal. Once the samples were trimmed, another quality report was conducted, through the FastQC tool, to ensure the quality of both the trimming process and samples.

Alignment

As for the alignment process, it can be separated in two steps, the genome indexing and mapping. For this process, there are several bioinformatics tools at our disposal. STAR (Spliced Transcripts Alignment to a Reference) is a recent software widely used for this step, and can resolve both the genome indexing and mapping, specially for accurate alignment of high-throughput of RNA-seq data [45], thus it was the software used for reads alignment (https://quay.io/repository/biocontainers/star, tag 2.7.8a–0). The genome indexing requires the complete genome of the species to which the samples belong to. Therefore, the *Rattus norvegicus* complete genome (mRatBN7.2.gff) was retrieved from the NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/GCF_015227675.2/). Once the reference genome is introduced, all the samples can be mapped with the target species genome, identifying introns, exons, among other features. The STAR software also allows the output of all the alignments for each sample sorted by coordinates of the genome in the BAM format to enable the next quality control check for mapping.

Before the quality check for mapping, the flagstat tool from SAMTools can be used to perform a quick analysis to the samples aligned BAM file to confirm the alignment was successfully executed. Similarly, the program featureCounts from the subread package (https://quay.io/repository/biocontainers/subread, tag 2.0.1–h5bf99c6_1) also offers a more detailed report for counting genomic features such as genes, exons, promoters, etc [46].

Finally, to end the quality control check for the mapping step, a more detailed analysis was conducted through the qualimap application (https://quay.io/repository/biocontainers/qualimap, tag 2.2.2d–1). Qualimap produces a report for each sample including the alignment rate, genomic origins of reads, transcript coverage profile and junction analysis. Alignment reads focus on the number of mapped reads, alignments, alignments to genes, to no features, among others; genomic origin gives a percentage of reads to its genomic origin, exonic, intronic, intergenic or overlapping; transcript coverage profile refers to which end 5' or 3' was most covered in the alignment. Ideally, the values in the 5'-3' bias should be 1, however it is known that the poly-A selection can lead to high expression in the 3' area rising the 3' bias of the sample [47]. Finally, the junction analysis is performed, which gives the total number of reads with splice junctions and the 10 most frequent junction rates. Splice junctions are reads that contain a former intron in a mature mRNA. Lastly, it also contains a section where it displays several graphical representations of the previous results [48].

After the second quality check, the amount of reports produced can make difficult the overall sample visualization, thus the use of the MultiQC tool (https://quay.io/repository/biocontainers/multiqc, tag 1.10–py_1) can provide a single report visualizing the output of multiple other tools across many samples in a quick and easy way, allowing the identification and proper correction of outliers and batch effects that can be missed in the early stages of analysis [49].

Afterwards, the read counts can be executed, through the python library HTSeq (https://quay.io/repository/biocontainers/htseq, tag 0.13.5–py39h70b41aa_1) which facilitates the rapid development of scripts for HTS (High Throughput Sequencing) data analysis and processing [50]. HTSeq can convert the resulting BAM files from mapping into text files, creating a matrix with two columns, the first with the Gene-ID obtained through the GFF file from the target species, in this case, *Rattus norvegicus*, and a second column with the raw counts of that gene in the specific sample, allowing a lighter file to contain all the necessary information to continue the pre-processing.

To finish the pre-processing, a merge of the samples HTSeq output would facilitate the further analysis of the samples. Therefore, a simple R script from Ahmed Alhendi (https://github.com/AAlhendi1707/htseq-merge/blob/master/htseq-merge_all.R, April 16, 2021) was used that allows the merge of each sample matrix into one single text file with the first column corresponding to the Gene-ID, and one column for each sample with the raw counts of the corresponding gene.

The remaining pre-processing was accomplished through R Studio an open-source interactive version for R (version 4.1.1) through a manufactured script for a complete analysis for all samples [51]. The overall matrix was then filtered to remove genes that were not expressed in every sample and reads that were not assigned to any gene. Additionally, metadata was added to the matrix, by naming the columns after the samples and creating groups of columns (samples) to identify the three different stages that were analysed, P1, P10 and P60. Afterwards, to perform normalization, and further pre-processing of the samples, the R package DESeq2 was used [52]. This method detects and corrects dispersion estimates that are too low through the dependence of dispersion modelling on the average expression strength in all samples. In summary, to each gene a GLM (generalized linear model) with a logarithmic link is applied, as follows:

$$log_2 q_{ij} = \sum\nolimits_r x_{jr}\, \beta_{ir}$$

where q corresponds to the mean quantity for a determined i gene, in j samples, considering its dispersion while also using a design matrix elements x and coefficients β. Additionally, it automatically detects outliers using Cook's distance and removes these genes from further analysis, thus completing the pre-processing for the bulk-RNAseq samples. An important side note is that DESeq2 normalization does not account for gene length so that the analysis remains unbiased [52].

2.1.2 Bulk RNA-seq data visualization and maturation

After normalization and filtering, the package DESeq2 also provides the tools for gene expression analysis allowing a comparative analysis between the samples. It uses shrinkage estimators for dispersion and fold change. Fold Change (FC) is a comparative measure that describes the quantitative change between a given group A compared to B. In this work, it is imperative to use Fold Change to compare samples, thus determining the differential expression between them. However, log2 (FC) will be used for a better approach to the biological perspective and better data visualization since the value 0 means both groups are equally expressing the gene, positive log2 (FC) means the gene is overexpressed in A, and consequently, negative log2 (FC) means it is overexpressed in B.

Before the differential expression analysis, from the normalized matrix counts, it was possible to first plot a PCA (Principal Component Analysis) for the samples. PCA is a common technique for increasing interpretability through the reduction of datasets dimensionality, which tries to minimize the loss of information. This analysis can create uncorrelated variables that successfully maximize variance. These new variables, principal components (PCs), are still defined by the data provided hence making PCA an adaptive data analysis technique [53]. The PCA plot allows a quick 2-dimension visualization of the data and how the samples differ from each other. Furthermore, heatmaps are also a common way to represent high throughput data, showing which genes are more common among the samples and how their expression varies among them.

DEA (Differential Expression Analysis) was then performed across the three stages creating three separate results, P1 versus P10, P1 versus P60, and P10 versus P60. The p-value measures the probability that a difference would occur by a random chance in the obtained results, Typically, results that have a p-value below 0.05 are considered to be statistically significant. However, if the data is a result of several comparisons, it is advised to use the p-adjusted value, e.g. through the Bonferroni correction, in which the p-values are multiplied by the number of comparisons, thus a p-adjusted value under 0.05

bring more statistical relevance to the analysed data. For each of the 3 resulted sets, the top 20 genes that were over-expressed (log2(FC)>= 2) and under-expressed (log2(FC)<= -2) were pulled for further analysis. As for the visualization of such results, three separate violin plots were created, one for each comparison, that allow to identify the genes that were more differently expressed between the two stages analysed.

After the DEA, GSEA (Gene Set Enrichment Analysis) (software version 4.1.0) [63,64] was performed to identify which pathways would be more expressed for each stage comparison. Through a normalized matrix of the dataset and an auxiliary file with the correspondent metadata, GSEA was performed. The Reactome pathway database was chosen to identify the pathway sets to which genes were involved in. The same three stages were compared as the gene expression comparisons (P1 versus P10, P1 versus P60 and P10 versus P60). For each comparison, the GSEA generated enrichment plots for both the compared stages, one report for each enrichment set found, a heatmap containing all the involved samples and the top 50 genes more enriched in each sample. Finally it also compiled two final reports, one for each stage containing their respective gene sets and the following values:

- Size, which represents the number of genes identified in the particular set;
- ES (Enrichment Score), which is calculated by first ranking the genes by their normalized expression levels and a running sum statistic is calculated from a prior defined gene signature present through the database, it is defined as the maximum deviation from zero of the running sum (Fig. 8) [54];

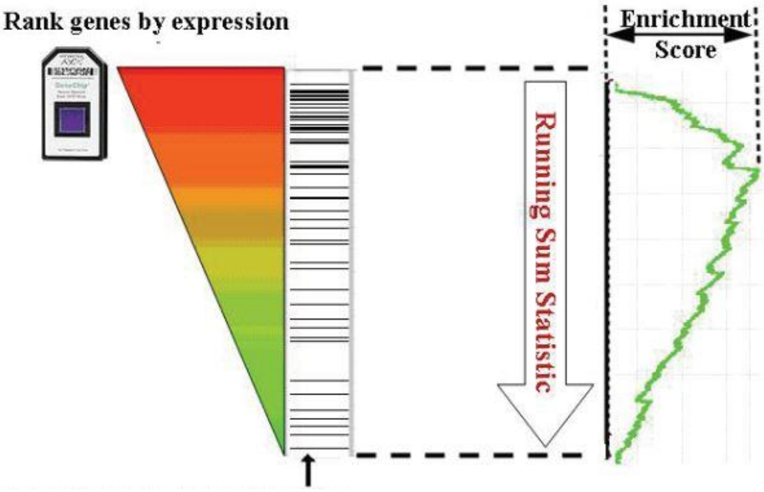

Figure 8 | *Enrichment Score calculation through ranking the genes by expression and comparing to a prior defined gene signature present in the database (adapted from* [54]*).*

- NES (Normalized Enrichment Score), which is the ES after it has been normalized across all the analysed sets;
- Nominal p-value, the statistical significance of the ES;

- FDR (False Discovery Rate) q-value, which stands for the estimate probability that the NES represents a false positive;
- FWER (Family Wise Error Rate) q-value, a more conservative approach that the NES represents a false positive;
- Rank at Max, the position in the ranked gene list at witch the maximum ES occurred.

It also contains the leading edge column, which displays three different statistics:

- Tags – the percentage of gene hits before the ES peak;
- List – the percentage of gene in the ranked list before the ES peak;
- Signal – the ES signal strength which combines the previous two statistics.

The leading-edge analysis was also performed after the GSEA, which allows a visual display of the overlap between the genes of the Reactome pathway sets.

Finally, the enrichment map visualization was performed from the GSEA results for each stage comparison, with the following cut-offs: p-value 0.001, FDR 0.05 and similarity with an overlap coefficient of 0.75. The enrichment map can be displayed through Cytoscape (version 3.8.2) with the application EnrichmentMap (version 3.3.3) concluding the analysis of the Bulk RNA-seq samples.

2.2 ScRNA-seq analysis

ScRNA-seq was performed on mice with similar timestamps as the bulk RNA-seq in rats, P3, P11 and P60, with n=1 each. The sample processing was performed by Ana Mendanha Falcão at ScilifeLab (Sweden). Single cells were further processed using 10x genomics technology.

The scRNA-seq analysis, followed a similar path to the bulk-RNAseq, however one must consider the UMI, which removes the amplification noise from PCR and biases. Afterwards, one can collapse the reads with the same UMI allowing the count of transcripts (Fig. 9).

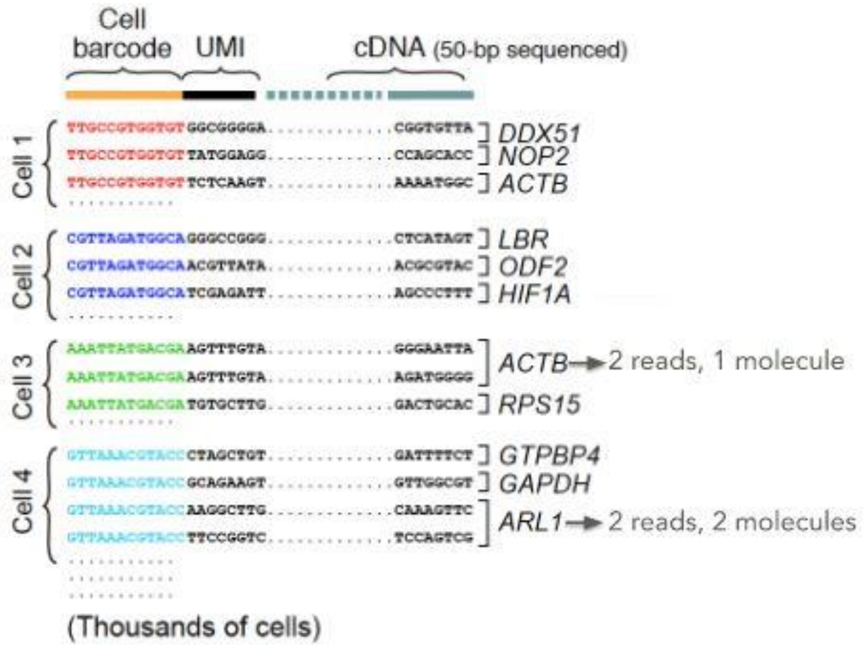
Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGT GGCGGGGA .... CGGTGTTA
TTGCCGTGGTGT TATGGAGG .... CCAGCACC
TTGCCGTGGTGT TCTCAAGT .... AAAATGGC
DDX51
NOP2
ACTB
Cell 2
CGTTAGATGGCA GGGCCGGG .... CTCATAGT
CGTTAGATGGCA ACGTTATA .... ACGCGTAC
CGTTAGATGGCA TCGAGATT .... AGCCCTTT
LBR
ODF2
HIF1A
Cell 3
AAATTATGACGA AGTTTGTA .... GGGAATTA
AAATTATGACGA AGTTTGTA .... AGATGGGG
AAATTATGACGA TGTGCTTG .... GACTGCAC
ACTB → 2 reads, 1 molecule
RPS15
Cell 4
GTTAAACGTACC CTAGCTGT .... GATTTTCT
GTTAAACGTACC GCAGAAGT .... GTTGGCGT
GTTAAACGTACC AAGGCTTG .... CAAAGTTC
GTTAAACGTACC TTCCGGTC .... TCCAGTCG
GTPBP4
GAPDH
ARL1 → 2 reads, 2 molecules
(Thousands of cells)

Figure 9 | *Example of a set of reads from scRNA-seq. After the alignment through the cell barcode, it is possible to collapse the reads that have the same UMI since it belongs to the same molecule, this allows the removal of amplification noise and further biases* [73].

2.2.1 ScRNA-seq data processing and analysis

The CellRanger pipeline v3.1.0 (from 10X genomics), with default settings, was used to demultiplex the output of Illumina bcl files, align reads to the mm10-3.0.0 reference genome and extract counts matrix for each sample. The count matrix of each sample was analysed independently first and then they were combined for better comparisons between different postnatal samples.

The package "Seurat" (Version 4.0.5) from R allows loading 10x data and later to create an object which allows the visualization, treatment, and analysis of the obtained sample matrixes [55]. After each sample was loaded, the correspondent metadata was added, including age (either P3, P11 or P60), tissue (all from lateral ventricle) and sex (all mixed). The first approach to scRNA-seq data is to ensure the quality of samples by checking several parameters, such as:

- Number of counts – number of reads per cell that the sample contains.
- Number of features – number of genes per cell in the sample.
- Percentage of mitochondrial RNA – percentage of mitochondrial RNA in each cell. Cells with an high percentage of mitochondrial RNA are associated with cell death [56].
- Percentage of ribosomal genes – percentage of ribosomal genes in each cell.
- Percentage of hemoglobin RNA – percentage of hemoglobin RNA in the cells. Red blood cells have no nuclei and can be present in the sample. A high percentage of hemoglobin

RNA indicates that the cell encapsulated for sequencing is a red blood cell and thus should be removed.

Initially, the quality control removed cells with less than 500 expressed genes, as well as genes expressed in less than 5 cells. Furthermore, cells with more than 25% of mitochondrial or hemoglobin reads were removed from the data as they are either not viable, or are red blood cells, respectively. Cell cycle scores calculation also needs to be considered, since cells can vary their transcriptome depending on the cell cycle stage. Typically, a high number of cells in the S or G2/M stage can weaken the sample, since their profiling is temporary dedicated to cell division rather than their phenotype, therefore making difficult to classify them. Cell cycle scoring calculation was performed through the bioMart package from BioConductor (version 2.50.0), which provides the database necessary to identify the genes that are associated with the cell cycles [57][58]. Doublets, droplets containing gene expression of two (or more) cells, were removed from the data due to the abnormal values for gene expression. Doublets can be predicted in each sample, according to the recovered number of cells from Illumina, through the tool DoubletFinder (version 2.0.3), which requires normalization and scaling (Fig. 10) [59].

We normalized the data of each sample separately by using LogNormalize function. In detail, feature counts for each cell were divided by the total counts, scaled by 1000 and natural-log transformed. Then, we scaled data by centred each feature's mean to zero and scaled by its standard deviation. DoubletFinder performs a pN-pK parameter sweep for all cells and after the visualization of the mean-variance normalized bimodality coefficient (BCmvn) score for each pK value, an optimal pK can be achieved by picking the highest point in the BCmvn distribution. Once the optimal pK value and doublets rate, which 10xGenomics provides, are found, DoubletFinder can successfully predict the doublets and further removal from the sample. After removing doublets, cells with more than 30,000 UMI counts and/or 6000 genes were also removed to avoid noise in downstream differential gene expression (DGE) analysis.

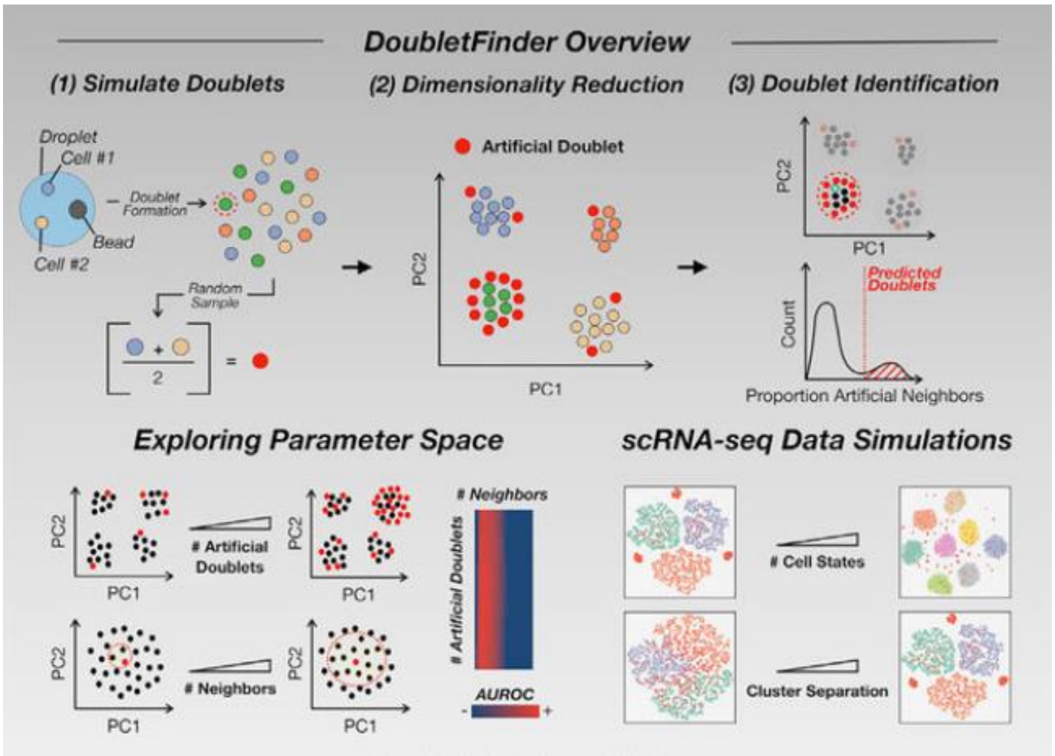

Figure 10 | *General overview of the DoubletFinder tool process of doublets removal. 1- Simulates the number of doublets based error rates. 2- Dimensionality reduction will help revealing doublet clusters. 3- Doublets identified based on the prediction (adapted from* [59]*).*

Dimensional reduction

After quality control, the expression data was normalized again. Due to the traits of the expression matrix, high dimensional and high sparse data, it is necessary to select a representative set of genes to estimate low dimension embeddings to reduce the noise in clustering. 2000 genes that exhibit high cell-to cell variation were selected by the 'vst' method in FindVariableFeatures function with default settings, followed by data scaling.

For the dimensionality reduction step, we started by performing PCA and converted data into 50 dimensional embeddings. The top 20 principal components, which explain most of the cell variation, were kept for future dimensionality reduction. However, PCA only interprets the linear regression, since scRNA-seq produces more complex/ dimensional data, it is advised to perform a second dimensionality reduction, often using tSNE or UMAP. T-SNE is an unsupervised, non-linear technique for exploration and high-dimensional visualization data. Essentially, t-SNE places the objects in a low dimensional space, while preserving neighbourhood identity [60], and then the T-distribution creates the probability distribution of points in the lower dimensional space. The deficiency of the t-SNE is that it only explains two-dimension variations. Since global structure is a key feature for the biological interpretation, we chose UMAP which interprets both non-linear variations and global relations to visualize the results.

To identify the clusters of cells, the graph based clustering approach Louvain clustering was used for all three samples. For Louvain clustering, the shared nearest neighbour (SNN) graph was built by estimating the neighbourhood overlap (Jaccard index) between every cell and grouping cells into communities by k-nearest-neighbours (k = 20). To determine the number of clusters, different values for Louvain distance were considered, 0.01, 0.05, 0.1, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9, independently. Usually, the higher resolution, the more clusters or subclusters are formed for the same data. However, these clusters might not have biological significance. Therefore, for this work the resolution was picked to identify the expected cell types in our samples, not focusing on subcellular types. Afterwards, the clustree package (version 0.4.3) can offer an intuitive overview of the clustering resolution increase process.

The K-means clustering algorithm was also considered as a possible approach, but this method focuses on finding groups that are not explicit labelled in the data. Since from our samples, it is deducible that several communities will form clusters according to their cell type, the Louvain algorithm was a more proper fit for the detection of such communities.

Finally, to end the individual sample analysis, cell type assignment is necessary. There are several predictive ways to assign cells to cell types and they all require a previous study or database to use as reference. Since there are too few studies for CP development, we combined canonical markers and top markers of clusters to define cell types. Several markers were took into account both from other studies such as M. Lehtinen [61], as well as the online database CellMarker, which can provide known cell type markers from previous studies [62]. After consulting the gene markers expression distribution, through the visual functions DotPlot and FeaturePlot, it is possible to correctly assign a cell type for each cluster represented.

Once cell type assignment was complete, all samples were merged to be compared, and allow more types of analysis. Initially, the samples were filtered for ribosomes, since we are now comparing a greater number of cells the amount of noise produced can increase and interfere with data analysis. Therefore, cells with ribosome gene percentage over 25% and cells expressing hemoglobin RNA were removed from the samples. Two types of integration were tested, Canonical Correlation Analysis (CCA) and Harmony [63], all after normalization, scaling and dimensionality reduction. Integration is a fundamental process for multiple scRNAseq datasets, CCA focuses on determining relationships between groups of variables in a dataset, it was used up to 30 dimensions and 2000 anchor features (number of features to start the anchor finding). While Harmony is a more complicated algorithm, scRNA-seq specific, that first soft clusters the sample, then finds centroids for each dataset, corrects dataset factors for each cluster and moves cells based on the soft cluster membership, it then iterates these four steps until cell cluster assignment is complete [63]. Harmony also used 2000 for the top variable features and ran up to 30 dimensions.

After integration, clustering was performed with an identical pipeline as for each sample, plots from t-SNE and UMAP for PCs 10, 15 and 20 followed by a K-nearest neighbour algorithm. Several resolutions were tested for the UMAP (0.05, 0.075, 0.1, 0.15, 0.2, 0.3) and the same cell markers were used for the cell type assignment.

After clustering and cell type assignment, the Differential Expression Analysis (DEA) was performed. The Seurat package allows to identify markers after clustering through the non-parametric Wilcoxon test. Therefore, several top markers were identified for each cell type, based on the adjusted P-value, and demonstrated through dotplots. Afterwards, the DEA was focused on cell types to identify the variance between age, similarly, the results can be visualized through barplots, taking into account the log2(FC) and adjusted p-value of the DEA results. Similar stage comparisons were performed for further comparison with the bulk RNA-seq analysis, P3 versus P11, P3 versus P60 and finally, P11 versus P60. Finally, the top 15 overexpressed features for each comparison were plotted.

2.3 Bulk RNA-seq and ScRNA-seq comparison

Due to the extensive output from the DEA analysis in scRNA-seq, the comparison between the two techniques was focused on the highest variance stage comparison. Furthermore, a deeper analysis was conducted to compare common genes from the most enriched cell type populations across the integrated samples in scRNA-seq compared with bulk data. Multiple genes were then analysed according to their ontology and both techniques were discussed through the relevance and reliability of their respective outputs.

3. Results and Discussion

In this section, the results from both bulk RNA-seq and scRNA-seq analyses will be displayed. Reports from bulk pre-processing are of difficult display since they are produced in HTML files and can be very extensive, thus most of pre-processing results will be on the supplementary section category.

However, the DEA and GSEA can easily be displayed through either violin plots, dotplots, barplots, or through software such as Cytoscape.

3.1 Bulk RNA-seq analysis

In the next section, a general overview of the bulk RNA-seq dataset analysis results is provided.

3.1.1 Pre-processing

Taking the initial fastqc reports and compiling them through multiqc allows a fast overview of the data's state. A high duplicate rate of reads was found in all samples (approximately 50%), as expected for RNA-seq analysis, the number of unique reads go from 7 million to 9 million.

As for the Phred score, it was already considered in good values ranging from 28 to 32. As expected, the per base sequence content shown a 3' bias, due to the selection of Poli-A in the sequencing process. Even though, the per sequence GC content was not shown as ideal, since most reads were in the range 40%-50% for each sample. Per base N content shown absence of base calls for the N, apart from two samples, however, both were lower than 0.2%. Overrepresented sequences also never cross the 1% line for each of the samples except for sample P10.2 which reached 1.13%, however it should not influence the downstream analysis.

All samples from P3 stage presented a significant amount of adapter content, however these sequences can be removed through trimming. After the trimming process, all adapter sequences were removed through Cutadapt. Most samples maintained 50% to 60% of duplicate reads, however Phred score improved across all samples, ranging from 32 to 36. Duplicate reads should not be filtered from RNA-seq samples without UMIs, since the removal of these sequences will also remove valid biological duplicates and most likely harm the analysis than provide benefits, even for paired-end data [64]. The GC content per sequence also improved across all samples, as well as the percentage of overrepresented sequences. Other parameters were maintained equal for the further analysis.

After using STAR to map the reads against the *Rattus norvegicus* genome from the NCBI database, the mapping quality review was performed using FeatureCounts and Qualimap, thus, the quantity of reads mapped, the percentage of assigned and aligned reads was obtained (Table 2).

Table 2 | *General overview of the mapping quality control obtained through MultiQC.*

Sample Name	*M Reads Mapped*	*% Assigned*	*M Assigned*	*% Aligned*	*M Aligned*
P1.1	21.3	47.8%	10.2	78.2%	13.7
P1.2	17.7	47.7%	8.4	78.0%	11.3
P1.3	21.7	42.1%	9.1	79.4%	14.0
P10.1	20.5	52.4%	10.7	83.7%	14.6
P10.2	23.1	65.6%	15.2	87.2%	18.3
P10.3	18.7	61.4%	11.5	85.1%	14.0
P60.1	20.5	60.3%	12.4	85.2%	15.3
P60.2	19.6	61.7%	12.1	86.4%	15.0

All the reads from the samples were successfully aligned over 75% to the *Rattus norvegicus* genome, and apart from the P1 samples, surpassed 50% of assigned reads. However, all samples obtained approximately 10 million assigned reads, which constitutes a reliable dataset for downstream analysis.

The full MultiQC reports for the quality control before trimming, after trimming and after mapping are included in the supplementary material in the sections A.1, A.2, A.3, respectively.

3.1.2 General Overview

After the quality control being complete, the first overall overview of the data was performed with PCA which allows a general overview of the samples and stages in a two-dimensional resolution (Fig.11).

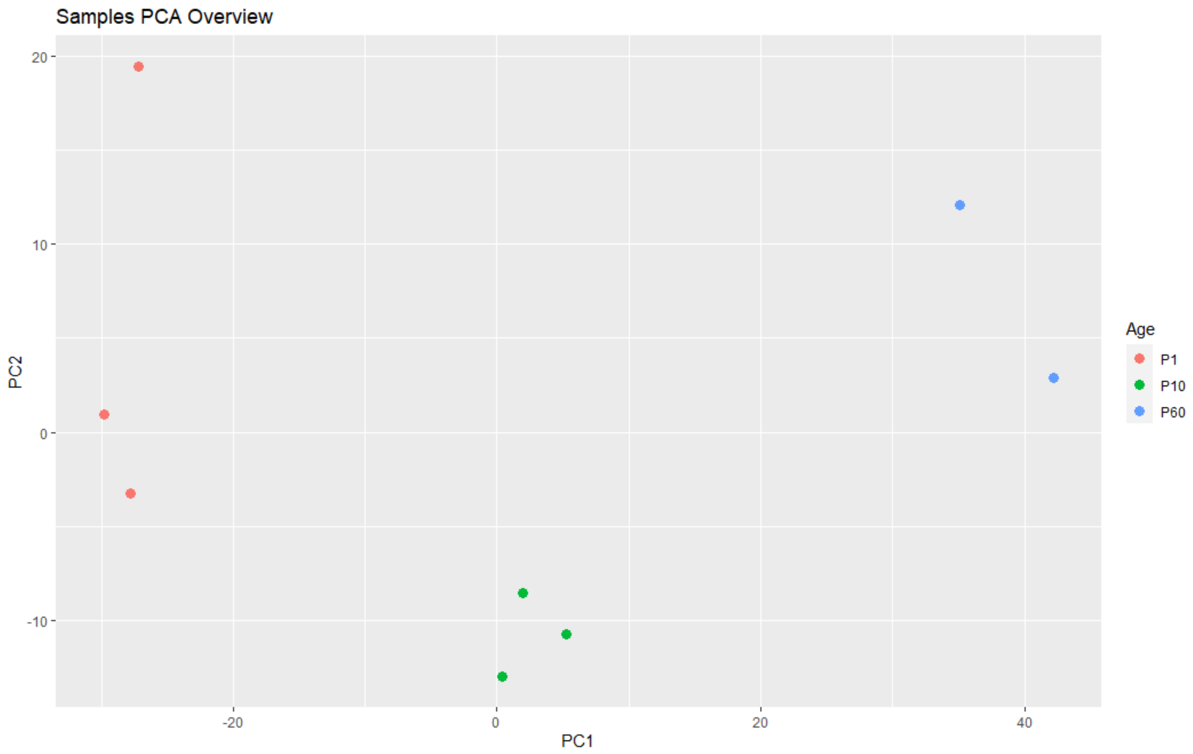

Figure 11 | *PCA plot of the rat CP transcriptome at 3 stages of development: P1, P10 and P60.*

As expected, samples with the same age demonstrate closer proximity than with different ages. Furthermore, the P10 samples also reveal a more neutral position when compared to P1 and P60 since they could be considered has the mid-developmental stage.

For the same purpose, an heatmap was also generated with the most 30 common genes and their variance across all samples after the VST (variance stabilizing transformation) being able to already identify possible markers related to the samples age (Fig.12).

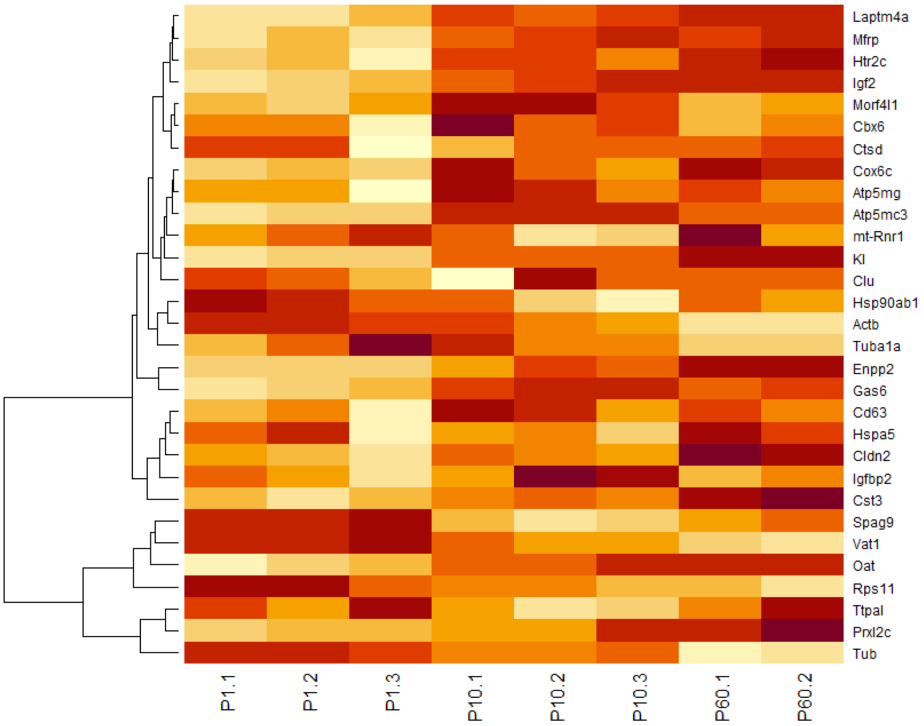

Figure 12 | *Heatmap overview with the most 30 top genes enriched in the different samples. A lighter colour reflects a more absent gene in the determined sample, whereas a darker/ stronger colour a more prevalent presence.*

Likewise, to the PCA, the P1 samples display a more distant phenotype to P60, when compared to P10, which reveals more intermediate and wider expression of genes at P10 stage. The genes *Kl* and *Enpp2*, involved in carbohydrate metabolic process and negative regulation of cell-matrix adhesion, respectively, were specifically enriched in P10 and P60 while *Actb* and *Tuba1a,* related with cell adhesion and mitotic cell cycle, respectively, were overexpressed in P1. The intermediate CP stage P10, displayed fewer enriched genes when compared to P1 and P60, for example, *Morf4l1* gene involved in DNA repair and regulation of cell growth.

3.1.3 Differential Expression Analysis

Regarding the DEA, the following comparisons between stages were performed: P1vsP10, P1vsP60, P10vsP60 which originated 3 main volcano plots displaying the genes enriched in each indicated time point.

Volcano plots are a simplistic visual display of DEA, since they can easily exhibit the designed thresholds, thus clarifying the biological perspective of its results.

The P1vsP10 DEA comparison aims to identify different developmental genes related to each stage (Fig.13).

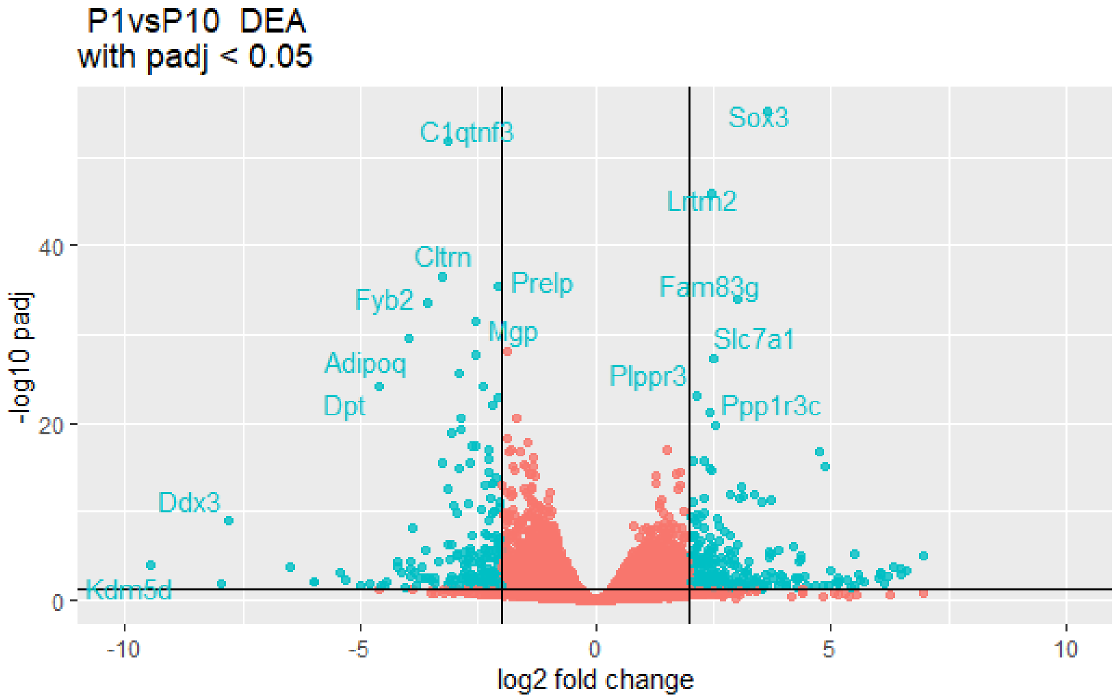

Figure 13 | *Volcano plot for the DEA between the ages P1 and P10. Thresholds for the significant FC are under -2 and over 2, while for the p-adjusted value is under 0,05. Genes included in these thresholds are represented in blue, while the others are red. More significant genes are also pointed in the map (low p-adjusted values and extremely lower/ higher FC).*

The P1 developmental stage was enriched in genes such as *Sox3,* which is a key transcription factor in developing and mature glia, and *Plppr3,* a phospholipid with signal transduction function, while at P10 the CP was enriched in the genes *Dpt*, *Cltrn,* which promotes collagen fibril organization and insulin secretion, respectively.

The comparison of CP at P1 versus P60 mice displayed the higher number of differentially expressed genes (Fig.14).

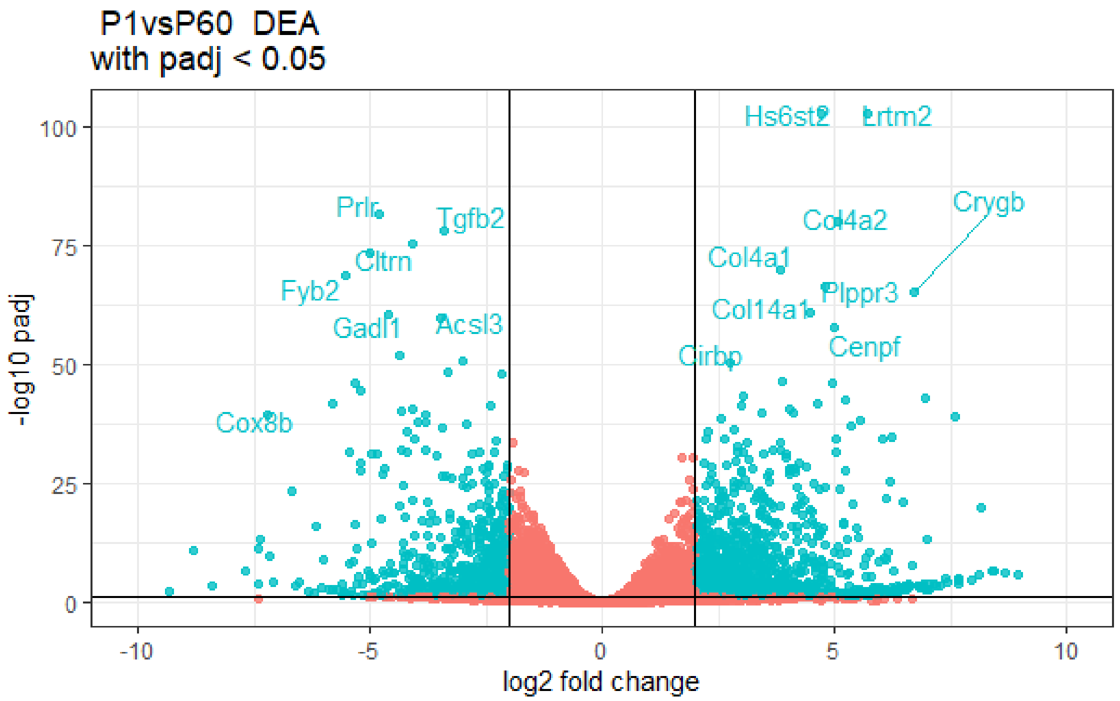

Figure 14 | *Volcano plot for the DEA between the ages P1 and P60. Thresholds for the significant FC are under -2 and over 2, while for the p-adjusted value is under 0,05. Genes included in these thresholds are represented in blue, while the others are red. More significant genes are also pointed in the map (low p-adjusted values and extremely lower/ higher FC).*

Over 6500 genes were identified as differentially expressed between the P1 and P60 stages. Almost 1300 genes were overexpressed on P1, such as, *Col4a1*, *Col4a2* and *Col14a1* which are all genes from the fibrillar collagen family, known to have a key role in the extracellular matrix organization and structure, and, therefore, in cellular adhesion in the early stages of development. In the P60 DEA, the overexpressed genes were *Prlr*, *Cox8b*, and *Acsl3*. *Prlr* is a receptor for the prolactin hormone, while *Cox8b* is a fundamental enzyme in the mitochondrial electron transport chain, involved in the energy metabolism and finally *Acsl3* is involved in both synthesis and degradation of cellular lipids. Overall, cellular matrix associated genes were enriched in P1 stage while metabolism associated genes were overexpressed in P60.

Lastly, comparing CP from P10 with P60 mice, unravelled similarities with the comparison between CP from P1 with P60 mice as expected, since both P1 and P10 are postnatal stages (Fig.15).

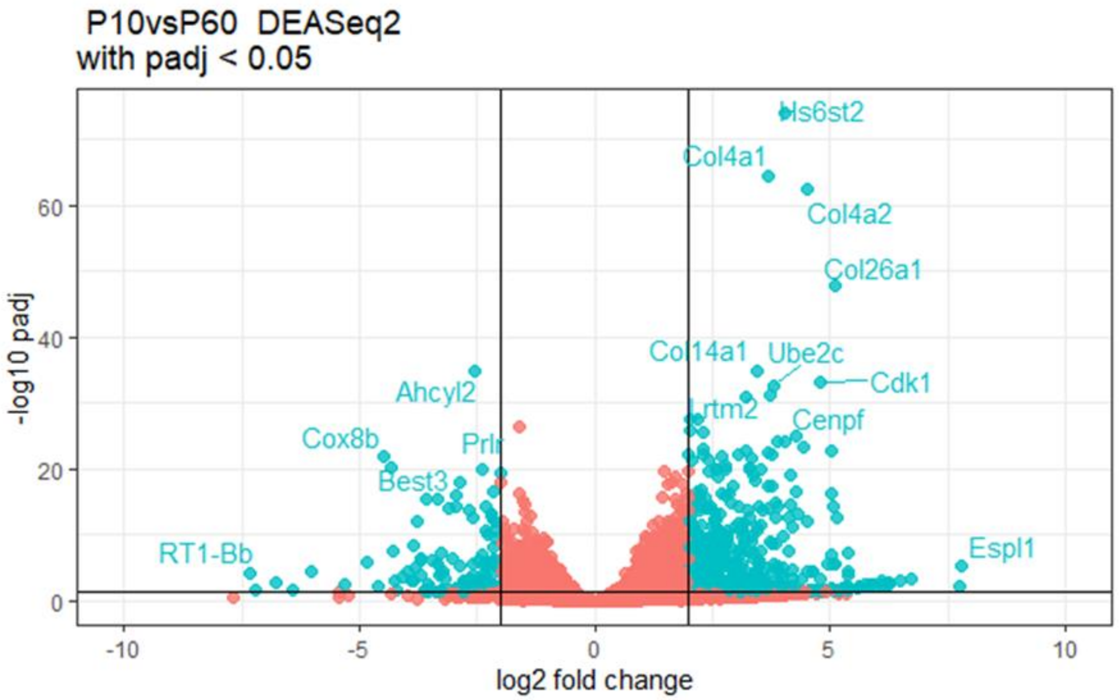

Figure 15 | *Volcano plot for the DEA between the ages P10 and P60. Thresholds for the significant FC are under -2 and over 2, while for the p-adjusted value is under 0,05. Genes included in these thresholds are represented in blue, while the others are red. More significant genes are also pointed in the map (low p-adjusted values and extremely lower/ higher FC).*

Likewise P1, the P10 DEA also shown enrichment of collagen family genes such as, *Col14a1*, *Col4a2* and *Col26a1* suggesting a tendency for cell adhesion and cellular matrix priority in early stages.

As for the DEA of this comparison in P60 also presented identical enriched genes such as *Cox8b* and *Acsl3*, as in the previous comparison against P1, however, besides immune associated genes like Cd74 (supplementary material section B.2) it also presented *Plp1* (supplementary material section B.3), a major gene for a myelin protein for the CNS, it has a high importance in both formation and maintenance of the multilamellar structure of myelin. These results suggest that P10 it is still not in a fully developed CP stage, and it mainly focuses on cell adhesion and proliferation, most likely in the epithelial cells.

DEA volcano plots with higher resolution can be consulted in the supplementary material for P1vsP10, P1vsP60 and P10vsP60 in the sections B.1, B.2 and B.3, respectively.

3.1.4 GSEA

P1vsP10

Following the DEA analysis, the GSEA was performed for the same comparisons as mentioned above to give insights about the pathways enriched for each sample, and their biological meaning. For all the stages, the complete reports in HTML format can be accessed in the supplementary material in the sections C.1, C.2 and C.3 for P1 versus P10, P1 versus P60 and P10 versus P60, respectively.

Out of 977 total gene sets found in both samples, 24 gene sets were significantly enriched in P1 (nominal p-value < 1%), while 112 were enriched in P10 (supplementary material section C.1). The main enriched gene sets in P1 were striated muscle contraction, DNA strand elongation and mitotic prometaphase, with the latest having a size of 193 genes involved. The gene set striated muscle contraction also achieved the highest NES with 2.06 with a size of 36 genes. The remaining gene sets are also mainly related with cell division and cycle. As for the P10 age, the main enriched gene sets were associated with the respiratory electron transport and citric acid TCA cycle, as well as pyruvate metabolism. Furthermore, several other gene sets are also related to energy metabolism. The most enriched gene set in P10 was the respiratory electron transport with NES correspondent to -2.91 and size 162. Overall, the NES was superior in P10 gene sets, as well as the number of their statistical significance.

As for the analysis of overlapped genes in the found gene-sets, on this comparison, it was only performed in P10 since P1 did not achieve a high number of significant gene sets needed for the overlap analysis. P10 displayed a high overlap of genes in the gene sets, as expected, since most are related to the energy metabolism and cell signalling.

As for the enrichment map visualization, it displays a more interactive version of the previous results (Fig. 16).

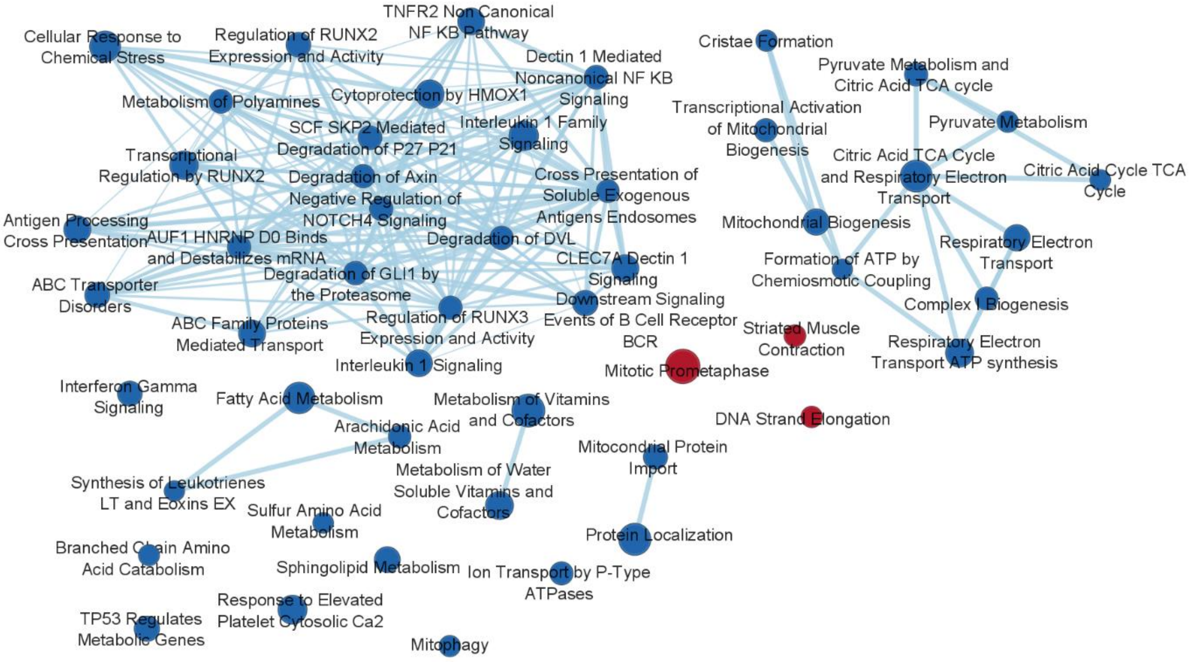
TNFR2 Non Canonical NF KB Pathway
Cellular Response to Chemical Stress
Regulation of RUNX2 Expression and Activity
Dectin 1 Mediated Noncanonical NF KB Signaling
Cristae Formation
Pyruvate Metabolism and Citric Acid TCA cycle
Metabolism of Polyamines
Cytoprotection by HMOX1
Transcriptional Activation of Mitochondrial Biogenesis
Pyruvate Metabolism
SCF SKP2 Mediated Degradation of P27 P21
Interleukin 1 Family Signaling
Transcriptional Regulation by RUNX2
Degradation of Axin
Cross Presentation of Soluble Exogenous Antigens Endosomes
Citric Acid TCA Cycle and Respiratory Electron Transport
Citric Acid Cycle TCA Cycle
Negative Regulation of NOTCH4 Signaling
Antigen Processing Cross Presentation
AUF1 HNRNP D0 Binds and Destabilizes mRNA
Degradation of DVL
Mitochondrial Biogenesis
Respiratory Electron Transport
ABC Transporter Disorders
Degradation of GLI1 by the Proteasome
CLEC7A Dectin 1 Signaling
Formation of ATP by Chemiosmotic Coupling
Downstream Signaling Events of B Cell Receptor BCR
Complex I Biogenesis
ABC Family Proteins Mediated Transport
Regulation of RUNX3 Expression and Activity
Striated Muscle Contraction
Respiratory Electron Transport ATP synthesis
Interleukin 1 Signaling
Mitotic Prometaphase
Interferon Gamma Signaling
Fatty Acid Metabolism
Metabolism of Vitamins and Cofactors
DNA Strand Elongation
Arachidonic Acid Metabolism
Mitocondrial Protein Import
Synthesis of Leukotrienes LT and Eoxins EX
Metabolism of Water Soluble Vitamins and Cofactors
Sulfur Amino Acid Metabolism
Protein Localization
Branched Chain Amino Acid Catabolism
Sphingolipid Metabolism
Ion Transport by P-Type ATPases
Response to Elevated Platelet Cytosolic Ca2
TP53 Regulates Metabolic Genes
Mitophagy

Figure 16 | Enrichment map of P1 versus P10 GSEA. The red nodes represent enriched gene sets in P1, while the blue nodes represent enriched P10 gene sets. Edges connecting nodes represent the overlap of genes between those gene sets.

As mentioned above, the enrichment map of the P1 versus P10, displays 2 main clusters (both enriched in P10), one referring to energy metabolism, which relates gene sets such as the complex I biogenesis, respiratory electron transport and the citric acid TCA cycle sharing a similar number of common genes as the leading analysis also proved. The second largest cluster with more connections refers to signalling pathways associated to the immune system, such as Interleukin 1 signalling, genes involved in cellular response to chemical stress, and others related to cell fate such as the NOTCH4 pathway. The P1 stage is also displayed with the top 3 gene sets mentioned above, although with no overlap genes, indicating a strong evidence of cell division priority at this age. Furthermore, it is also notable the existence of small clusters associated with fatty acid metabolism, vitamin metabolism and protein localization, enriched in P10, as well as other gene sets which are not directly related to the main 2 clusters such as, interferon gamma signalling, mitophagy and ion transport through P-type ATPases, which are indirectly related to the cell signalling and energy metabolism, respectively.

P1 versus P60

Similar to the previous GSEA analysis, the comparison between the P1 and P60 stages also displayed cell division/ cycle related gene sets enriched in P1. However, 84 gene sets were significantly enriched in P1 (nominal p-value < 1%) for this comparison, while 71 were enriched in P60 (nominal p-value < 1%) (supplementary material section C.2). The most significant enriched gene sets in P1 were the activation of the pre replicative complex with 2.36 NES and size 32, the resolution of D-loop structures, involved in cell repair mechanisms, and as in the previous GSEA analysis, the mitotic prometaphase. As for the P60 enrichment analysis when compared to P1, similar to P10, it also expresses gene sets related to the energy metabolism like the respiratory electron transport with -2.80 NES and size of 162 genes.

P1 shown high overlap of genes in cell division gene sets, as well as cell adhesion gene sets, as expected from the DEA analysis, while P60 also displayed a high overlap of genes in the energy metabolism gene sets mentioned above.

In the enrichment map analysis, although using the same parameters, it is observable a more significant enrichment in P1 gene sets when compared to the previous section, P1 versus P10 (Fig. 17). Besides a higher overall correlation between sets, there is also a prevalence of P1's clusters rather than P60's.

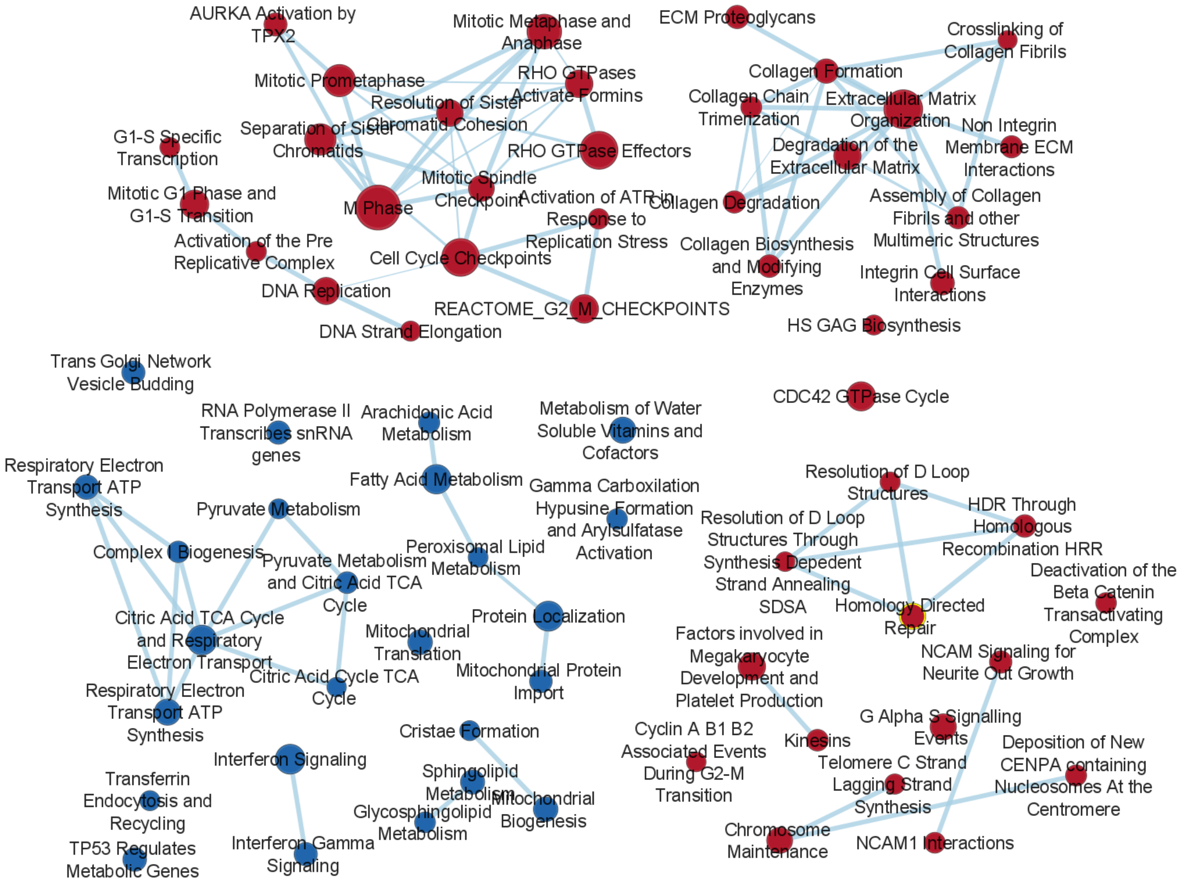
AURKA Activation by TPX2
Mitotic Metaphase and Anaphase
ECM Proteoglycans
Crosslinking of Collagen Fibrils
Mitotic Prometaphase
RHO GTPases Activate Formins
Resolution of Sister Chromatid Cohesion
Collagen Formation
Collagen Chain Trimerization
Extracellular Matrix Organization
Non Integrin Membrane ECM Interactions
G1-S Specific Transcription
Separation of Sister Chromatids
RHO GTPase Effectors
Degradation of the Extracellular Matrix
Mitotic Spindle Checkpoint
Mitotic G1 Phase and G1-S Transition
M Phase
Activation of ATR in Response to Replication Stress
Collagen Degradation
Assembly of Collagen Fibrils and other Multimeric Structures
Activation of the Pre Replicative Complex
Cell Cycle Checkpoints
Collagen Biosynthesis and Modifying Enzymes
Integrin Cell Surface Interactions
DNA Replication
REACTOME_G2_M_CHECKPOINTS
HS GAG Biosynthesis
DNA Strand Elongation
Trans Golgi Network Vesicle Budding
RNA Polymerase II Transcribes snRNA genes
Arachidonic Acid Metabolism
Metabolism of Water Soluble Vitamins and Cofactors
CDC42 GTPase Cycle
Respiratory Electron Transport ATP Synthesis
Fatty Acid Metabolism
Gamma Carboxilation Hypusine Formation and Arylsulfatase Activation
Resolution of D Loop Structures
Pyruvate Metabolism
Resolution of D Loop Structures Through Synthesis Dependent Strand Annealing SDSA
HDR Through Homologous Recombination HRR
Complex I Biogenesis
Pyruvate Metabolism and Citric Acid TCA Cycle
Peroxisomal Lipid Metabolism
Deactivation of the Beta Catenin Transactivating Complex
Citric Acid TCA Cycle and Respiratory Electron Transport
Mitochondrial Translation
Protein Localization
Homology Directed Repair
Factors involved in Megakaryocyte Development and Platelet Production
NCAM Signaling for Neurite Out Growth
Respiratory Electron Transport ATP Synthesis
Citric Acid Cycle TCA Cycle
Mitochondrial Protein Import
G Alpha S Signalling Events
Cristae Formation
Cyclin A B1 B2 Associated Events During G2-M Transition
Kinesins
Deposition of New CENPA containing Nucleosomes At the Centromere
Interferon Signaling
Telomere C Strand Lagging Strand Synthesis
Transferrin Endocytosis and Recycling
Sphingolipid Metabolism
Mitochondrial Biogenesis
Glycosphingolipid Metabolism
TP53 Regulates Metabolic Genes
Interferon Gamma Signaling
Chromosome Maintenance
NCAM1 Interactions

Figure 17 | Enrichment map of P1 versus P60 GSEA. The red nodes represent enriched gene sets in P1, while the blue nodes represent enriched P60 gene sets. Edges connecting nodes represent the overlap of genes between those gene sets.

As mentioned above, the enrichment map for the P1- P60 GSEA comparison, the stage P60 displays a major cluster related to the energy metabolism, as well as immune cell signalling sets such as interferon gamma and lipid and vitamin metabolism, similar to P10 in the previous analysis. However, P1 displays more prevalence in cell division/ cycle and DNA repair gene sets and a new cell adhesion cluster, involving genes sets that include the previous mentioned Collagen family genes in the DEA analysis, essential for the extracellular matrix organization.

P10 versus P60

The enrichment map analysis of CP from P10 versus P60 mice revealed a total of 95 gene sets significantly enriched in P10 (nominal p-value < 1%) and 28 significantly enriched in P60 (nominal p-value < 1%). Both P1 and P60, when compared to P10, display a low level of enriched gene sets, suggesting a higher prevalence of this developmental stage in phenotype variance (supplementary material section C.3).

Similar to the previous comparison, P10 also displays a high content of gene sets related to the extracellular matrix organization (like P1, in P1 versus P60), namely the collagen biosynthesis and degradation with 2.61 NES and a 66 gene size. Unlike P1, it also focuses on the degradation of collagen and extracellular matrix, suggesting a regulation process of this feature. Furthermore, P10 also displays enriched gene sets linked to cell division, such as mitotic spindle checkpoint and mitotic prometaphase. Interestingly, NCAM 1 interactions genes are also enriched in P10, the protein encoded by this gene is involved in cell-cell interaction and cell- matrix interactions [65], essential processes for cell adhesion and signalling, further strengthening the hypothesis of epithelial cells maturation at this particular stage.

As for the P60 enriched gene sets, these are also identical to the P1 versus P60 analysis, mostly immune cell signalling and energy metabolism, being the interferon signalling the top enriched gene set with a NES of -2.20 and a size of 73 genes. However, there are more uncommon gene sets such as transferrin endocytosis and recycling, and particularly, gene sets linked to insulin, reported in the literature as the CP releases insulin via serotonin signalling [66].

Regarding the enrichment map analysis, it displays 4 main clusters, three belonging to P10 and one to P60 (Fig. 18).

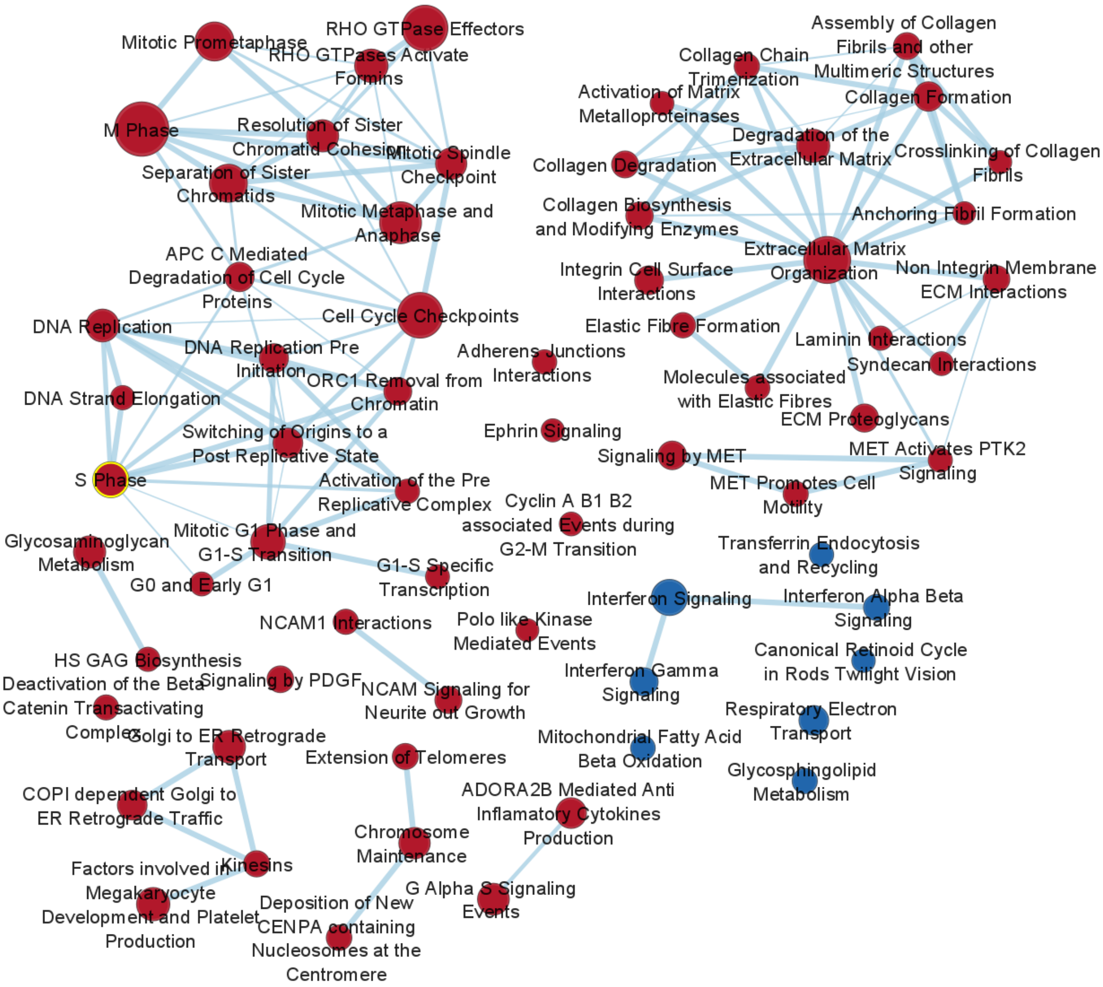
Mitotic Prometaphase
RHO GTPase Effectors
RHO GTPases Activate Formins
Assembly of Collagen Fibrils and other Multimeric Structures
Collagen Chain Trimerization
Activation of Matrix Metalloproteinases
Collagen Formation
M Phase
Resolution of Sister Chromatid Cohesion
Mitotic Spindle Checkpoint
Separation of Sister Chromatids
Degradation of the Extracellular Matrix
Crosslinking of Collagen Fibrils
Collagen Degradation
Collagen Biosynthesis and Modifying Enzymes
Mitotic Metaphase and Anaphase
Anchoring Fibril Formation
APC C Mediated Degradation of Cell Cycle Proteins
Extracellular Matrix Organization
Integrin Cell Surface Interactions
Non Integrin Membrane ECM Interactions
DNA Replication
Cell Cycle Checkpoints
Elastic Fibre Formation
DNA Replication Pre Initiation
Adherens Junctions Interactions
Laminin Interactions
ORC1 Removal from Chromatin
Syndecan Interactions
DNA Strand Elongation
Molecules associated with Elastic Fibres
ECM Proteoglycans
Switching of Origins to a Post Replicative State
Ephrin Signaling
Signaling by MET
MET Activates PTK2 Signaling
S Phase
Activation of the Pre Replicative Complex
MET Promotes Cell Motility
Cyclin A B1 B2 associated Events during G2-M Transition
Glycosaminoglycan Metabolism
Mitotic G1 Phase and G1-S Transition
Transferrin Endocytosis and Recycling
G0 and Early G1
G1-S Specific Transcription
Interferon Signaling
Interferon Alpha Beta Signaling
NCAM1 Interactions
Polo like Kinase Mediated Events
Canonical Retinoid Cycle in Rods Twilight Vision
HS GAG Biosynthesis
Deactivation of the Beta Catenin Transactivating Complex
Signaling by PDGF
NCAM Signaling for Neurite out Growth
Interferon Gamma Signaling
Respiratory Electron Transport
Golgi to ER Retrograde Transport
Mitochondrial Fatty Acid Beta Oxidation
Extension of Telomeres
Glycosphingolipid Metabolism
COPI dependent Golgi to ER Retrograde Traffic
ADORA2B Mediated Anti Inflamatory Cytokines Production
Chromosome Maintenance
Factors involved in Megakaryocyte Development and Platelet Production
Kinesins
G Alpha S Signaling Events
Deposition of New CENPA containing Nucleosomes at the Centromere

Figure 18 | Enrichment map of P10 versus P60 GSEA. The red nodes represent enriched gene sets in P10, while the blue nodes represent enriched P60 gene sets. Edges connecting nodes represent the overlap of genes between those gene sets.

P10 enriched clusters comprise extracellular matrix organization (top right), cell division and cycle (top left) and several cell signalling pathways, either through NCAM1, PDGF or ephrin. P60 enriched clusters comprises immune cell signalling via interferon and energy metabolism such as the respiratory electron transport gene set.

Overall, the GSEA indicates a strong phenotype variance in P10, when compared to the remaining stages. It also reveals a prevalence of cell division and cycle, extracellular matrix organization gene sets for the early age stages P1 and P10, and energy metabolism, immune cell signalling gene sets for later age stages P10 and P60.

3.2 Single-cell RNA-seq analysis

In this section, the scRNA-seq results will be displayed. Initially, the samples were filtered and analysed separately until the cell type assignment process, and afterwards were merged, integrated, and DEA was performed for further comparison.

3.2.1 Filtering process

The results from the filtering process can be observed for each sample individually. Plots regarding the raw data, cell cycle scores and doublet removal process can be found in section E of the supplementary material.

CP from P3

The CP from P3 was obtained from 10x genomics protocols contained a total of 18599 different genes and over 8000 cells. After the removal of cells with less than 500 different genes and genes expressed in less than 5 cells, despite the few cells presenting more than 25% mitochondrial/ hemoglobin genes these were also removed. The cell cycle scores for the phases G2M and S were also close to 0, therefore no filtering was necessary for these features. The doublet removal performed through DoubletFinder successfully predicted 460 doublets, however, to remove further noise from unpredicted doublets, high gene expression cells were also removed, ending with a total of 17450 genes and 6178 total cells (Fig. 19).

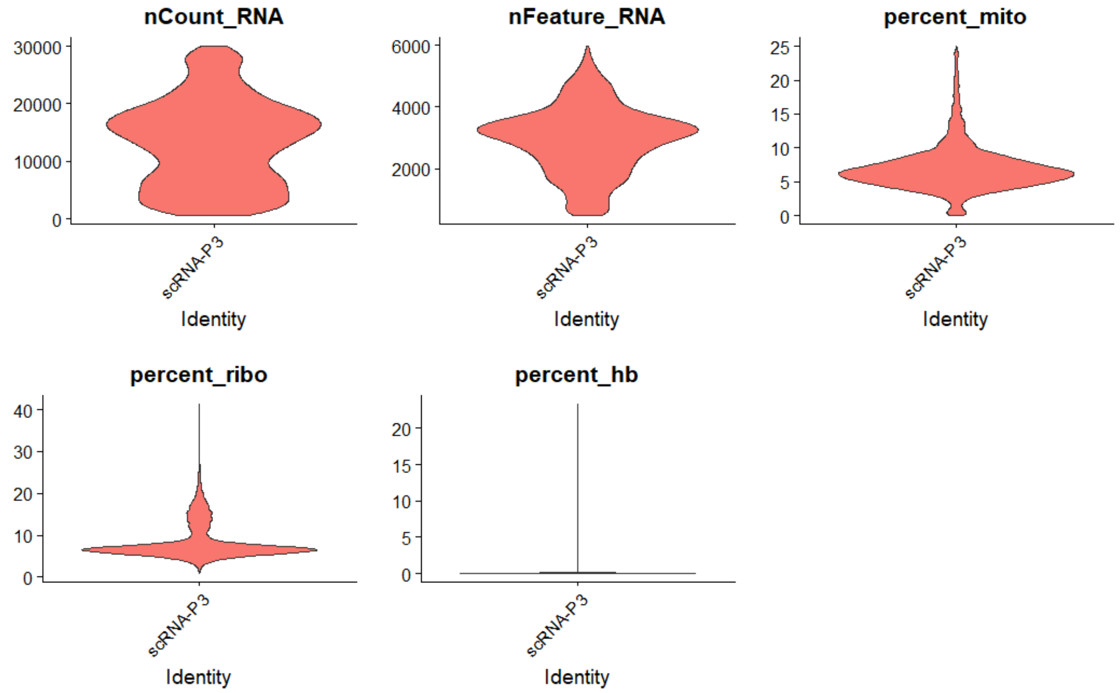

Figure 19 | General features of the P3 sample after filtering. Counts represent reads per cell while features represent genes per cell. Three percentages are associated to QC, percentage of mitochondrial/ ribosomal/ hemoglobin genes per cell.

Overall, each cell contains on average 15000 reads and 3000 genes, which is enough for cell assignment and DEA. Both ribosome genes and mitochondrial genes percentage are low indicating a good cell viability in the filtered sample. There are no red blood cells since the percentage of hemoglobin genes is practically non-existent.

CP from P11

The P11 sample contained the most cells from all samples, over 17000 and 31000 different genes. However, it also contained cells expressing a high percentage of mitochondrial genes, which suggests a higher percentage of cells with low viability, which were removed from the analysis. As for the remaining percentage of ribosomes and hemoglobin, it shown relative low values. After the cutdown of cells expressing high expression of mitochondrial, ribosomes and hemoglobin genes, the sample remained with around 17000 genes and 14000 cells. Due to the higher cell quantity, around 2000 doublets were predicted which were later removed remaining 12000 cells. After further removal of cells expressing a high quantity of genes similar to the previous sample, the filtered P11 sample contained 10922 cells and 17660 different genes (Fig. 20).

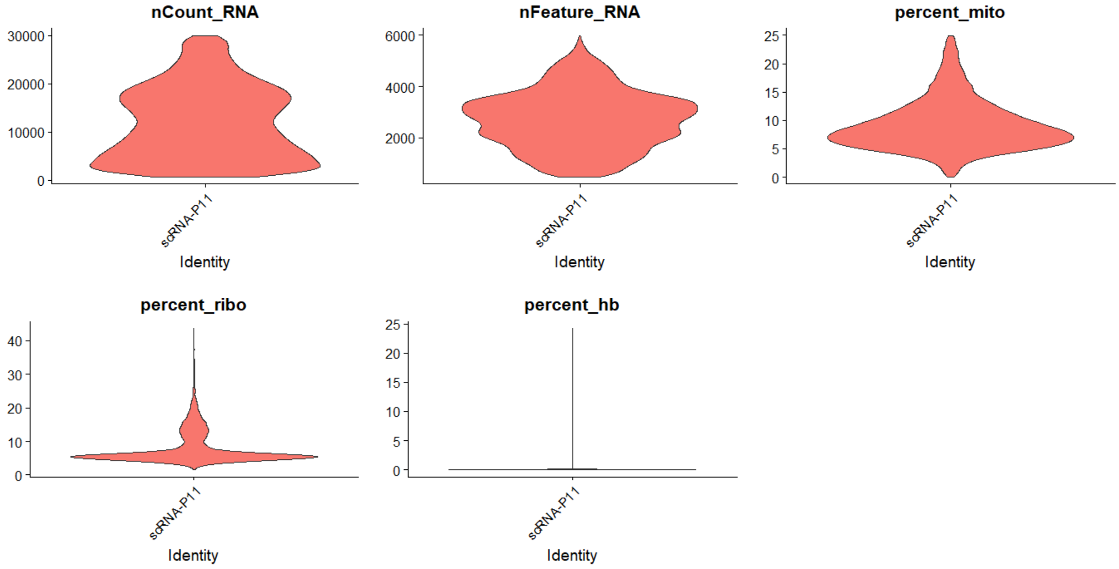

Figure 20 | General features of the P11 sample after filtering. Counts represent reads per cell while features represent genes per cell. Three percentages are associated to QC, percentage of mitochondrial/ ribosomal/ hemoglobin genes per cell.

Most of the cells contained around 3000 different genes and low mitochondrial/ ribosomal gene expression. Despite the higher number of dead cells, the P11 sample after filtering still have 11000 cells for downstream analysis.

CP from P60

The P60 sample contained the least number of cells and genes across all samples, with only 5678 and almost 17000, respectively. Furthermore, it also contained a high quantity of cells expressing a high percentage of mitochondrial (up to 75%) and ribosomal genes (up to 45%). However, the percentage of hemoglobin genes was non-existent. Due to the low number of cells, after the filtering, 147 doublets were predicted and further removed, and after the final removal of high gene expression cells, the filtered sample remained with 2919 cells and almost 16000 different genes (Fig.21).

Despite the low number of cells, the quality proved to be good for downstream analysis, since most cells express 5000 different genes with more than 10000 reads.

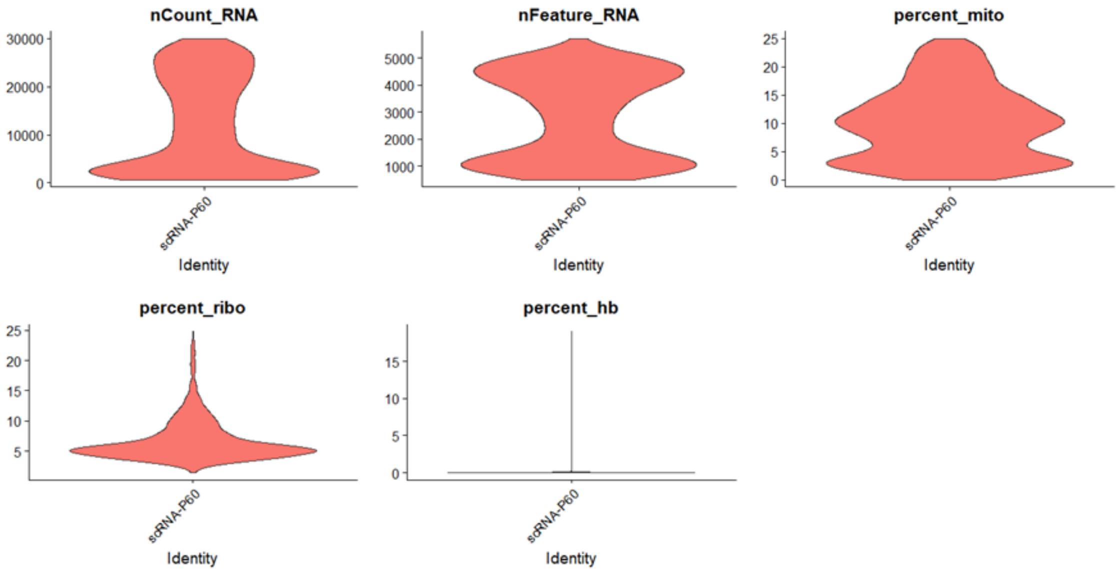

Figure 21 | General features of the P60 sample after filtering. Counts represent reads per cell while features represent genes per cell. Three percentages are associated to QC, percentage of mitochondrial/ ribosomal/ hemoglobin genes per cell.

3.2.2 Clustering and Cell type assignment

In this section, the results from the clustering and cell type assignment process can be observed for each sample individually. Plots regarding the dimensionality reduction through UMAP and T-SNE can be found in section F of supplementary material. Additionally, several resolutions for the Louvain and K-means algorithms in the clustering section can also be found in section G, including the cluster tree regarding the Louvain algorithm.

CP from P3 Sample

Through the analysis of PCs weights, it was revealed that 15 or 20 PCs were sufficient to explain the variance in the dataset as such we have used 20PCs to continue the clustering analysis (section F.1 in supplementary material).

The Louvain algorithm also allowed a more consensual overview of clusters rather than the K-means algorithm, thus it was the chosen algorithm for downstream analysis. The clustering resolution was set at 0.05 as it allowed to identify the main different cell types in the CP. Increasing the clustering resolution could identify possible subclusters of cell types in the sample (section G.1 in supplementary material).

Regarding cell type assignment, canonical markers were used to identify the expected populations: epithelial, mesenchymal, immune, and endothelial cells. This initial approach allowed the easy identification of the revealed cell types, while further analysis allowed the identification of a subcluster of epithelial cells, responsible for the formation of cilia, which is to be expected in early stages of development such as P3, a process known as ciliogenesis (Fig. 22) [67].

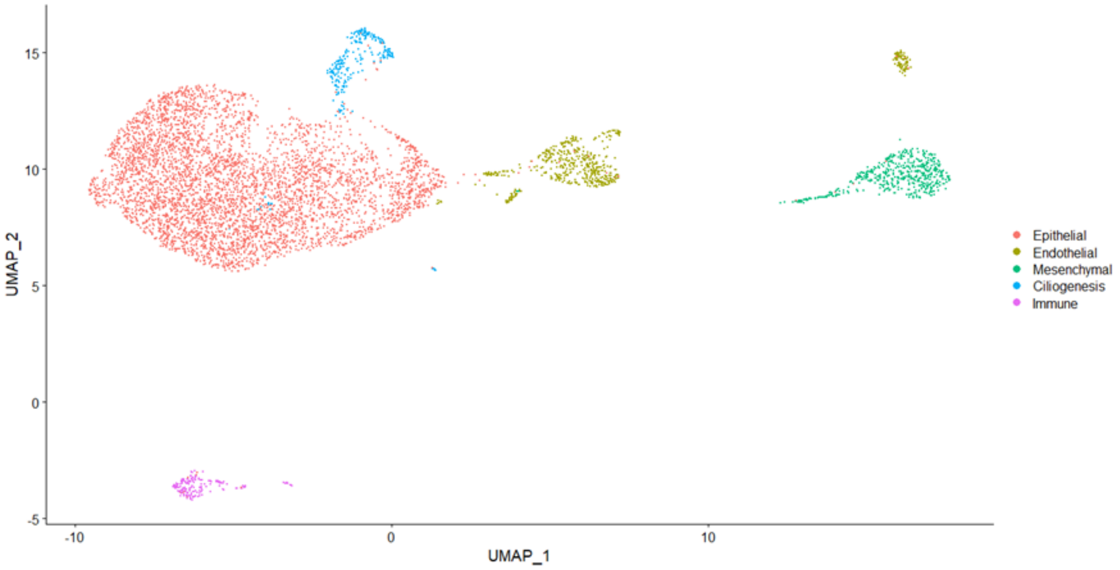

Figure 22 | *UMAP of the CP from P3 after cell type assignment.*

The usage of canonical markers not only allowed the identification of the expected cell types, but also the identification of the ciliogenesis subcluster. The used markers can be consulted in Figure 23 for each cell type population.

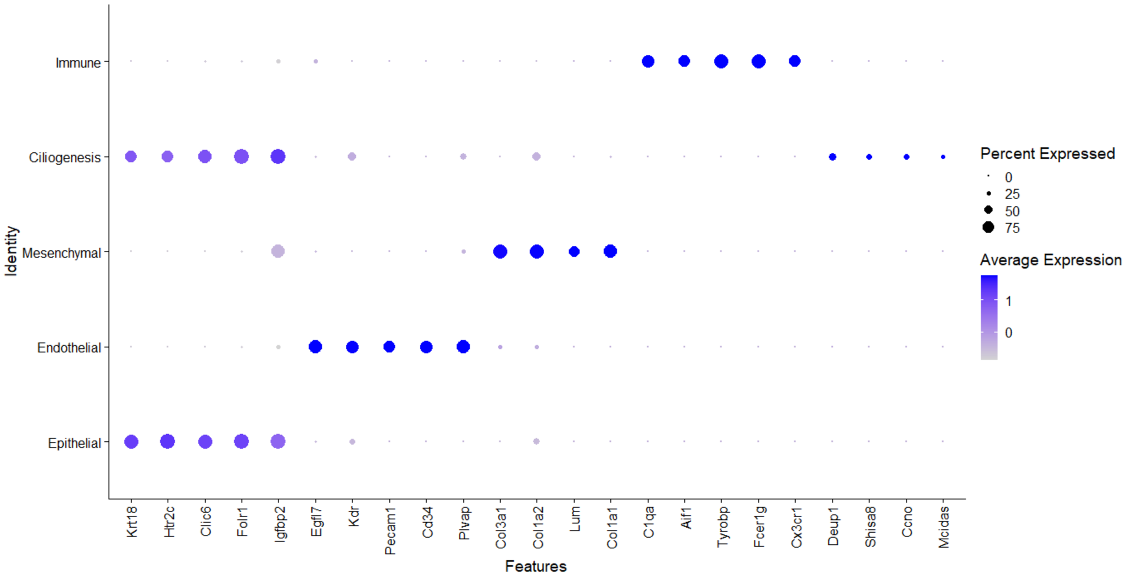

Figure 23 | *Dotplot of the used canonical markers for cell type assignment in the P3 sample.*

Despite the identification of these cell types, others, such as neurons, were not identified in this sample. Oligodendrocytes were also in low number and thus not defining a cluster. Of note, increasing the number of samples (and thus cells) processed would allow to increase clustering resolution leading to the identification of more cell clusters, such as distinct immune cell types, and neuroglia populations.

CP from P11

The clustering process in the P11 sample displayed a similar curve for standard deviation in the number of dimensions. Likewise in the previous sample, clustering was performed using 20PCs in the P11 sample. However, both 10 and 15 dimensions (on UMAP) could also be considered due to the visual identification of clusters (section F.2 in supplementary material).

The Louvain algorithm was the most effective clustering option than the K-means since the spatial position of clusters assignment were in agreement with the expected cell types from the CP (section G.2 in supplementary material). The Louvain algorithm with the lowest resolution was used for further downstream analysis. An increased resolution could also be a good option for further identification of subclusters within cell types, nevertheless, the number of samples would have to be increased.

Regarding the cell type assignment, canonical markers were used to identify cell types. Of notice, a new distinct population arose at the developmental stage P11, neuroglia (Fig. 24). Although the ciliogenesis markers seemed to be expressed, it was not possible to assign a cluster due to insufficient number of expressing cells. However, it is possible that through an increase in resolution, and mostly, by increasing the number of samples analysed, the epithelial subcluster of ciliogenesis would be identified. Furthermore, the immune cells also displayed several small clusters due to higher variance of cells, when

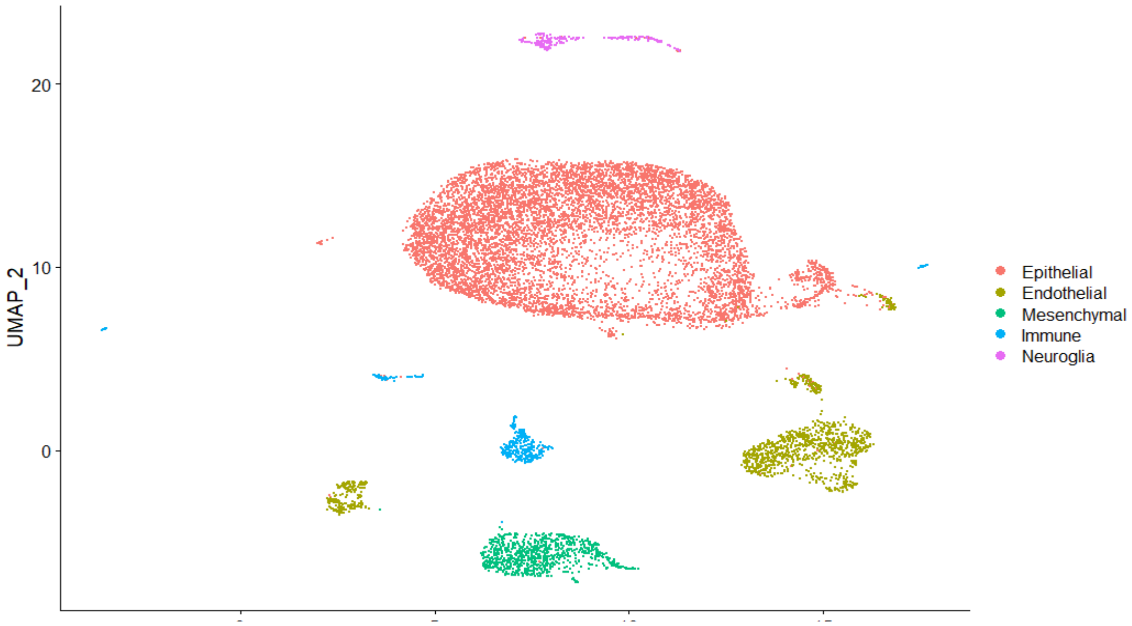

Figure 24 | *Overview of the 20PC UMAP of the sample P11 after cell type assignment.*

compared to the other cell types, increasing resolution and sample size, could potentially reveal the different types of immune cells in the sample, such as, macrophages, T-cells, and NK cells.

The used canonical markers for CP cell types were used to identify of cell types according to the clustering process (Fig.25).

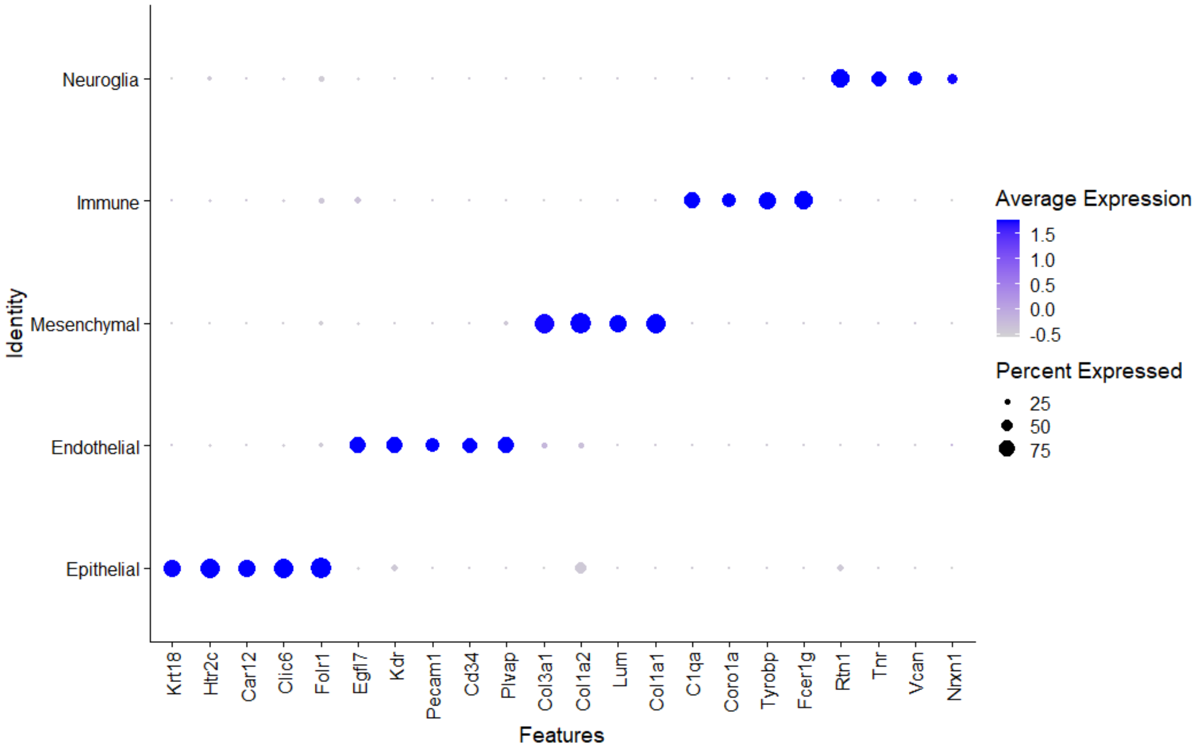

Figure 25 | *Dotplot of the used canonical markers for cell type assignment in the P11 sample.*

CP from P60

Similarly to the other samples, the standard deviation curve for PCs weights has also shown a rapid decrease. In order to maintain an identical approach towards the samples, the clustering was performed with 20 dimensions, however, there aren't significant differences in the visual plots with different dimensions in either the UMAP or T-SNE reduction approaches (section F.3 in the supplementary material).

Regarding clustering algorithm, likewise in other samples, the Louvain distance maintained a more solid approach towards the biological perspective of the clustering process (section G.3 in the supplementary material). Identical to the sample P3 the same resolution was opted which allowed the further identification of the expected cell types despite the considerable difference in the number of cells (approximately 50%).

The cell type assignment process used the same canonical markers as the previous samples, however only the epithelial, endothelial, mesenchymal, and immune cells were found in the sample (Fig. 26).

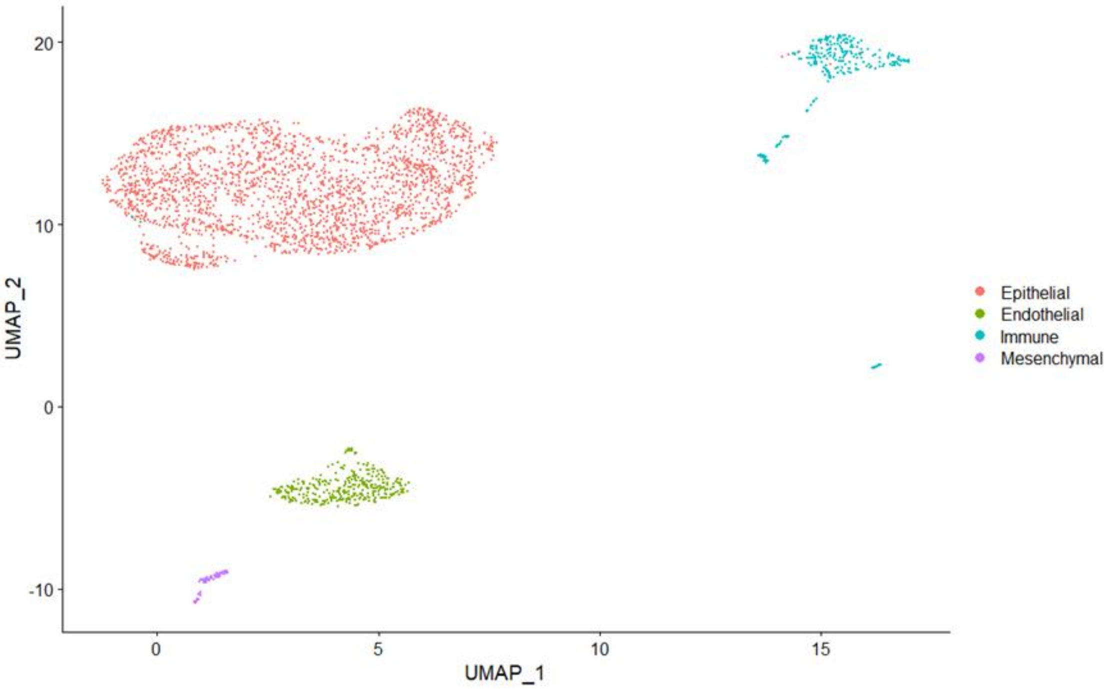

***Figure 26** | Overview of the 20PC UMAP of the sample P60 after cell type assignment.*

As expected in the adult stage, the ciliogenesis markers were not expressed in the sample since they are related with a transient stage of maturation of epithelial cells occurring at early stages in development. Furthermore, neuroglia cells were also in a very small numbers, not enough to define a visible cluster. The immune cells cluster also displayed several apparent subclusters as the P11 sample, suggesting a wider variety of cell type, although in minor quantity. The cell type assignment also proved to be efficient with the usage of the same cell type markers (Fig.27).

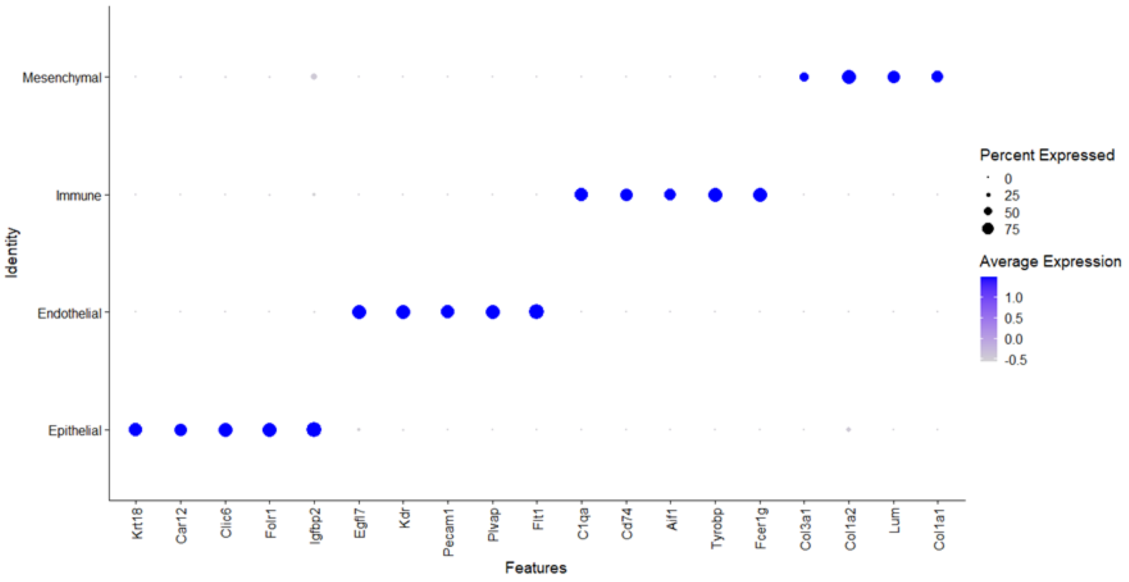

Figure 27 | *Dotplot of the used canonical markers for cell type assignment in the P60 sample.*

3.2.3 Integration

Since the individual samples were analysed previous to integration, the filtered versions were merged, scaled and normalized. In order to visualise the batch effects from the samples, a comparison between both integration algorithms and a non integrated version was performed (Fig. 28).

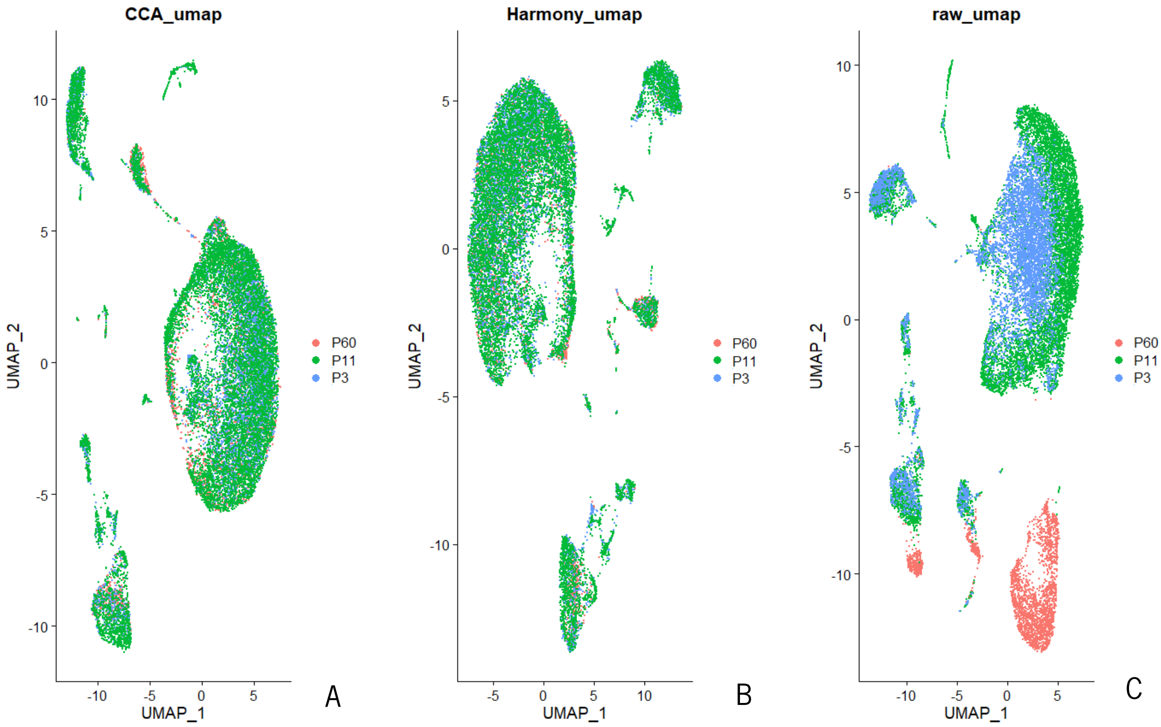

Figure 28 | *UMAP comparison between the two methods for integration CCA and harmony, and the merged samples. A – 15PC UMAP of the merged samples after integration through the CCA algorithm; B - 15PC UMAP of the merged samples after integration through the Harmony algorithm; C - 15PC UMAP of the merged samples.*

Both algorithms seemed to remove the batch effect successfully when compared to the raw merged sample UMAP (Fig. 27-C). However, due to a slightly clearer definition of clusters in the CCA integration, it was the chosen method for downstream analysis. Similar to the individual analysis of each sample, it is visible after integration the discrepancy between the number of cells in each sample where P11 holds more than the two other samples combined (P3 – 6152, P11 – 10 836, P60 – 2816).

The dimensionality reduction performed under 10, 15 and 20 PCs can be consulted in the supplementary material section F.1 alongside the tested resolutions for clustering through the Louvain distance algorithm (section F.2). The lowest tested resolution was picked since it would allow the identification of the expected cell types, parallel to the individual samples analysis.

The cell type assignment process through the used canonical markers successfully identified the same populations as in the individual analysis (section F.3 in supplementary material) allowing a general overview of the merged dataset (Fig. 29).

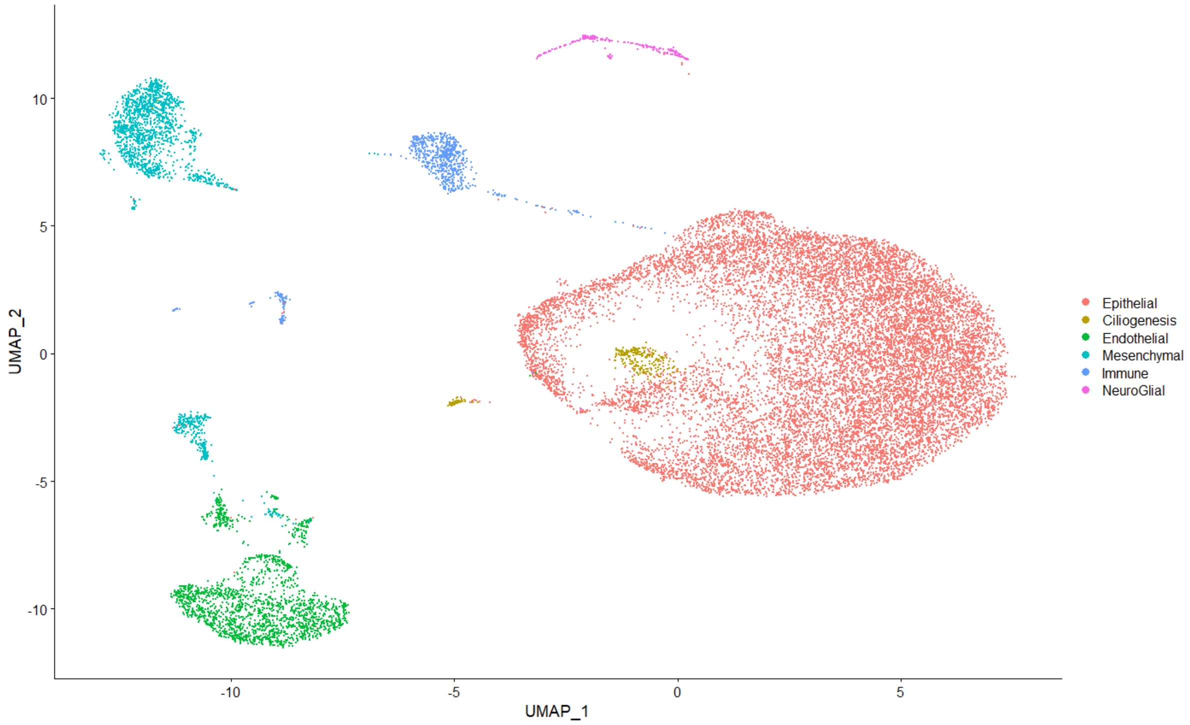

Figure 29 | *General UMAP overview of the merged samples after integration, clustering, and cell type assignment.*

The UMAP plot of the merged samples identified the ciliogenesis subcluster despite the low percentage of cells compared to the other cell types clusters, while multiple subclusters could be identified through an increased resolution in the mesenchymal and immune populations, that can emerge due an increase in the number of cells analysed.

Before the DEA analysis between the timepoints, the top 5 markers for each cluster were also plotted to possibly identify new markers for the correspondent cell types or reinforce the used ones (Fig. 30).

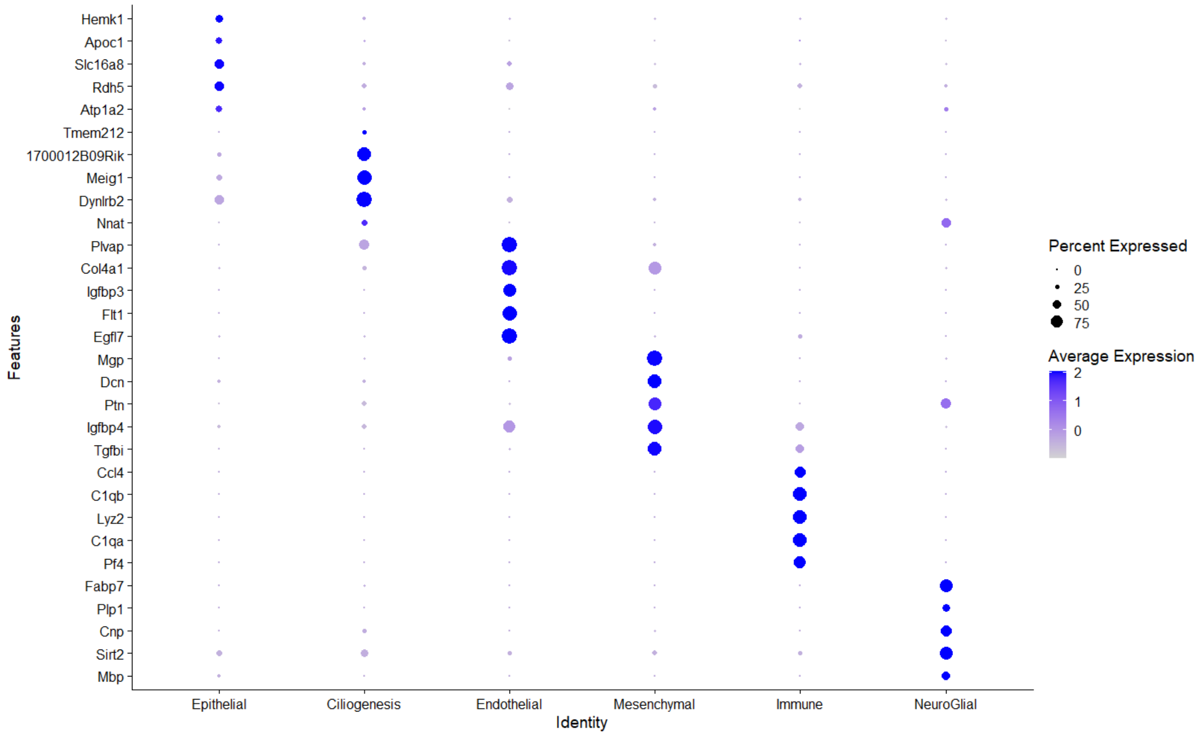

***Figure 30** | Dotplot of the top 5 expressed genes for each cell type across all samples.*

From the analysis of the top 5 gene markers for each cell type, only two genes were canonical used for the cell assignment process. The C1qa gene marker for immune cells is involved in the C1 antigen-antibody complex activation, that initiates the classical complement system, responsible for the ability of antibodies and phagocytic cells to clear either microbes or damaged cells, belonging to the innate immune system. The Plvap gene encodes an endothelial cell-specific membrane protein responsible for microvascular permeability and is also one of the canonical markers displayed in figure 30.

Despite the remaining canonical markers not being observable in the top 5, does not mean they are not enriched in their respective cell type (Section F.3 in supplementary material). Certain demonstrated genes could also potential become new biological markers, such as, *Egfl7*, responsible for promoting endothelial cell adhesion, *Mgp*, which is associated with the organic matrix of bones and cartilages, and finally, *Fabp7*, known to be associated with neurogenesis since it is required for the radial glial fiber system in the developing brain, a necessary system for the migration of immature neurons [68].

3.2.4 Differential Expression Analysis

DEA was then performed on similar stages as the bulk RNA-seq for each major cell type encountered, meaning the subcluster ciliogenesis was unavailable for DEA analysis due to insufficient cells in the P60 stage, which is to be expected since ciliogenesis is a typical process of developmental phases whereas P60 represents the adult stage, similarly, the neuroglial cluster was also specifically encountered in P11, but not on the other time points, thus the DEA wasn't performed as well.

Due to the high throughput of the DEA, this section will only cover the three stage comparisons in epithelial cells, since it was the most enriched population type across all samples, however it should be considered that we have only a n=1 per stage which means results should be considered preliminary and interpretated with caution. Furthermore, the different number of cells in each stage (P3 – 4732, P11 – 7845, P60 – 2138) can also affect the DEA. The remaining plots associated with the populations of mesenchymal, endothelial, and immune cells can be consulted in the supplementary material sections H.1, H.2 and H.3, respectively.

P3 versus P11

The comparison between the stages P3 and P11 is expected to contain less differences in gene expression than the other comparisons due to both being postnatal stages (Fig. 31).

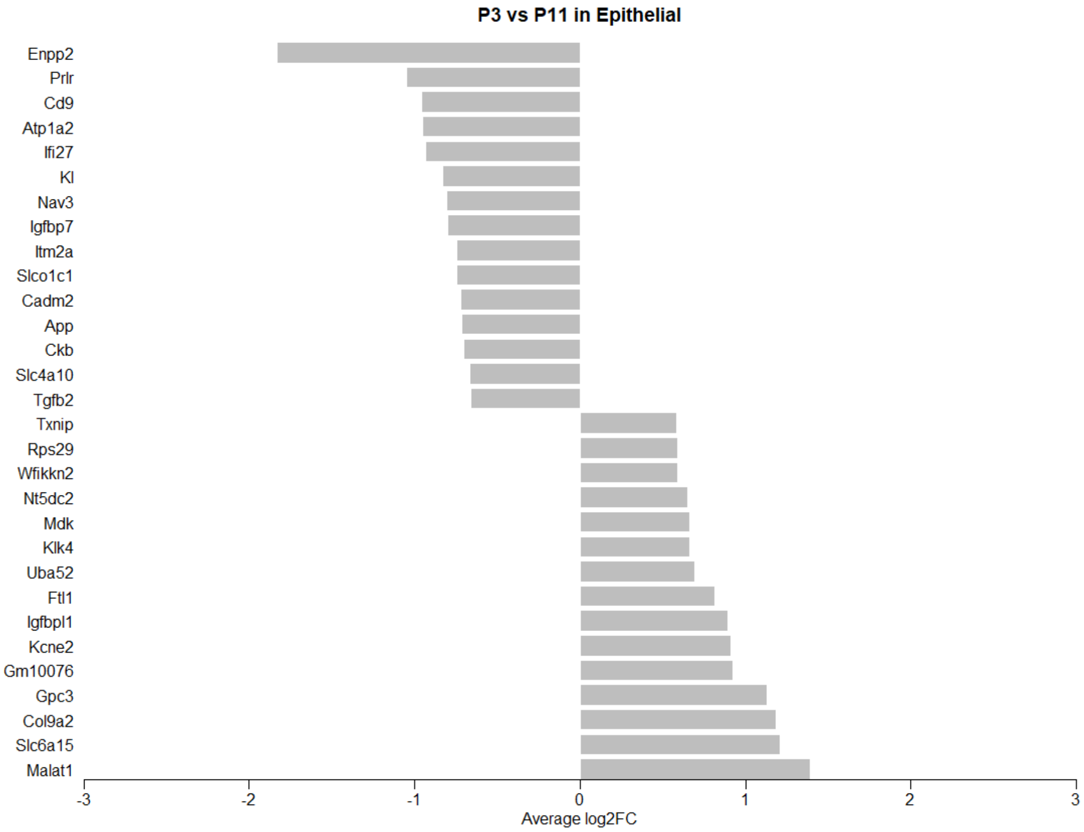

Figure 31 | *Barplot of the DEA results between P3 and P11. All the represented genes have P-adjusted value under 0.05.*

The top 15 overexpressed genes in P3 and P11 do not cross the log2(FC) mark of absolute 2, this could be due to the reasons mentioned above regarding the number of cells and samples that can impact the power of the analysis. Genes such as *Cd9* and *Gpc3* enriched in both stages and are associated with the regulation of cell proliferation. Furthermore, *Klf4*, a known TF involved in the embryonic development is enriched in P3. *Prlr*, a prolactin hormone receptor, was enriched in P11, and is associated with epithelial cell differentiation [69], which can be associated with later stages of development such as P11.

P3 versus P60

The comparison between the P3 stage and adult shows the most significant differential expression between all, despite the low number of epithelial cells present in P60 (Fig. 32).

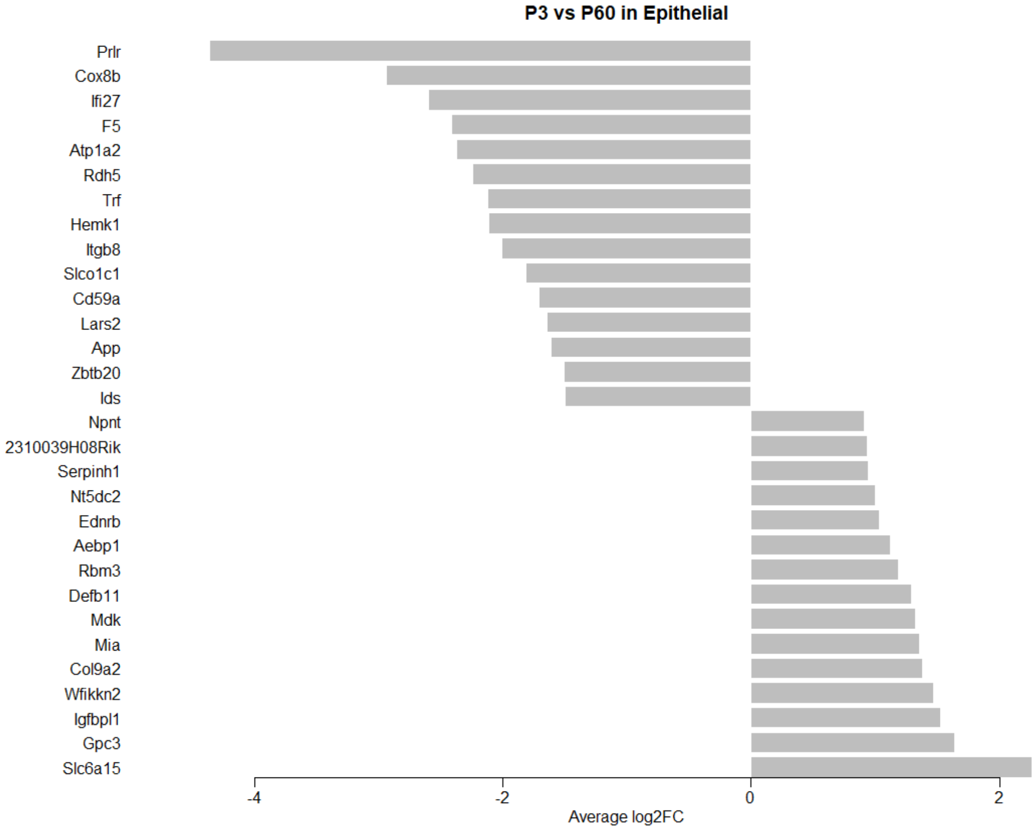

Figure 32 | *Barplot of the DEA results between P3 and P60. All the represented genes have P-adjusted value under 0.05.*

From this stage comparison, similar genes have shown to be overexpressed in P3 as in the previous, such as *Mdk*, *Gpc3*, suggesting a distinct phenotype in P3 against the other stages despite the low average gene expression FC. Furthermore, the *Prlr* gene expression also suggests to increase with age since its more overexpressed in P60 than P11 (when both compared against P3). The P60 overexpressed gene

Cox8b, has also been shown to be related to energy metabolism, which is to be expected in the adult stage when compared to P3.

P11 versus P60

As for the last comparison, it should be identical to the previous (P3 versus P60) due to the similarities shown between the P3 and P11 comparison since they are both developmental stages (Fig.33).

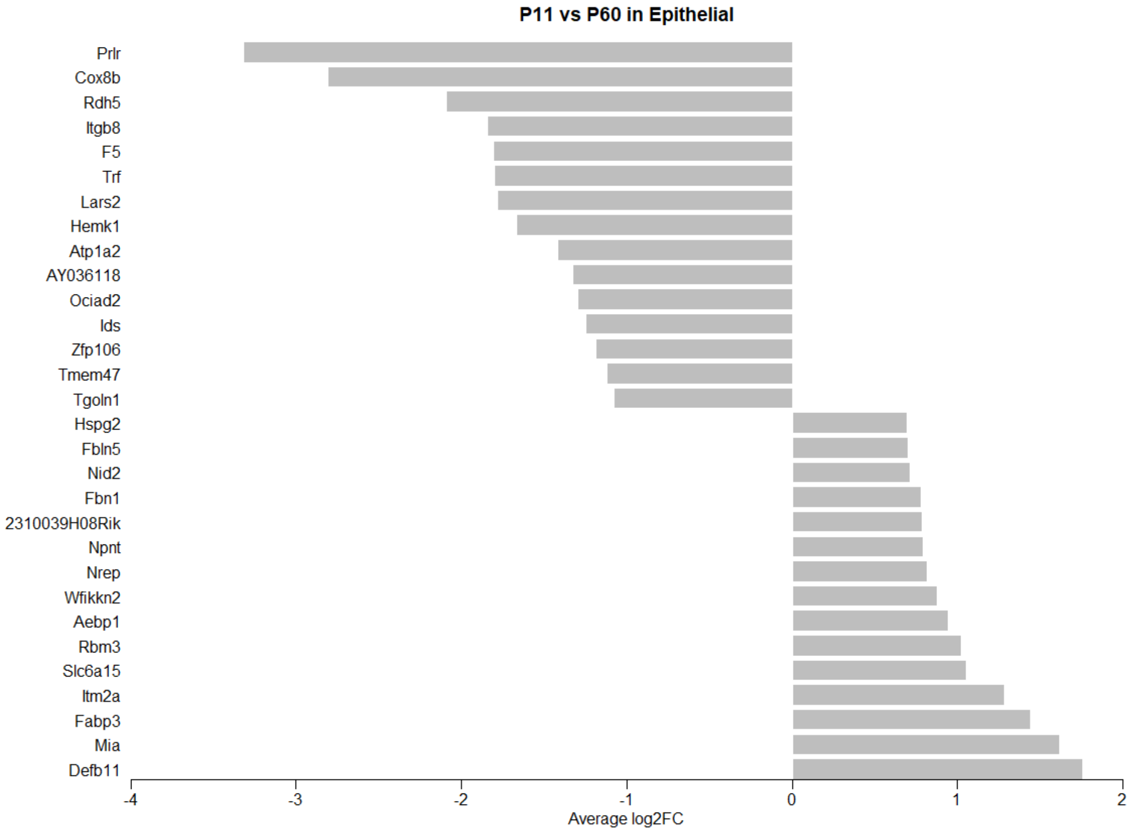

Figure 33 | *Barplot of the DEA results between P11 and P60. All the represented genes have P-adjusted value under 0.05.*

Likewise, the current comparison shown similar overexpressed genes in P60, while genes such as Mia, Aebp1, Rbm3 are also overexpressed in P3 against P60 which suggests that can be potential age markers, specifically associated with developmental stage of the CP.

3.3 Comparative analysis between bulk and Single-cell RNA-seq

As two techniques for transcriptomic analysis were performed, a comparison between the two outputs and their similarities should be made. Due to the extensive output from single-cell technique, this comparison was focused on the DEA between the stages P3 (P1 in bulk) and P60 to identify common genes that were found enriched in each cell stage. Despite the range difference in the number of genes, which is significantly higher in bulk, and the different organisms used in both techniques (*Mus musculus* in scRNA-seq and *Rattus norvegicus* in bulk RNA-seq), it is expected that at least some ortholog genes should similarly enriched in the different CP stages. Since the GSEA analysis wasn't performed in scRNA-seq this section will exclusively be focused on the DEA between the stages P3 and P60.

Between both techniques, a total of 114 genes were encountered in both samples DEA output, however not all had significant expression on the scRNA-seq dataset (log2(FC)<1 or log2(FC)>-1). Overall, one particular gene was found to be overexpressed in P3 across all cell types and in bulk, *Marcksl1*, which is a gene involved in actin filament binding, associated with the development of the CNS and positive regulation of cell proliferation [70]. In the next sections, the cell type populations will be individually compared to the bulk RNA-seq output.

3.3.1 Epithelial Cells

From the DEA analysis in epithelial cells, a total of 11 genes (absolute log2(FC)>1) were found in common with the bulk analysis, and only 2 of them were found overexpressed in P3, the remaining 9 were overexpressed in P60 (Table 3).

Table 3 | *Comparison between the common genes found in the analysis of Bulk RNA-seq and epithelial cells in single-cell RNA-seq. All the genes represented have P-adjusted value under 0.05.*

Genes	scRNA-seq (Log2(FC))	Bulk RNA-seq (Log2(FC))
Col9a2	1.382	3.918
Igfbpl1	1.524	2.173
Prcd	-1.029	-2.873
Acsl6	-1.038	-3.054
Abat	-1.121	-2.248
Ltc4s	-1.239	-4.928

Slco1c1	-1.818	-2.443
Itgb8	-2.014	-4.066
F5	-2.418	-4.967
Cox8b	-2.940	-7.191
Prlr	-4.363	-4.802

The 2 genes overexpressed in P3 are *Igfbpl1*, associated with cell growth, and *Marcksl*. As for the overexpressed in P60 genes, are associated with several roles, such as lipid metabolism (*Ltc4s*, *Acsl6*), insulin production (*Igfbpl1*, *Abat*), as well as the previously mentioned *Prlr* and *Cox8b*.

The epithelial cells displayed the highest number of genes in common with bulk, possibly due to higher number of cells when compared to the other cell types.

3.3.2 Mesenchymal Cells

As for the mesenchymal cells a total of 10 genes were equally found in both techniques output, however, unlike the epithelial cells which had more overexpressed genes in P60, the mesenchymal cells have 7 genes overexpressed in P3 and 3 for P60 (Table 4).

Table 4 | *Comparison between the common genes found in the analysis of Bulk RNA-seq and mesenchymal cells in single-cell RNA-seq. All the genes represented have P-adjusted value under 0.05.*

Genes	scRNA-seq (Log2(FC))	Bulk RNA-seq (Log2(FC))
Col26a1	2.082	4.973
Mfap2	2.071	3.367
Col9a2	1.728	3.918
Marcksl1	1.572	4.290
Basp1	1.378	2.357
Col4a1	1.149	3.830
Ppic	1.118	2.025
Prelp	-1.058	-2.052
Inmt	-1.727	-4.966
Coch	-3.717	-6.675

The gene *Inmt* is associated with amine compounds metabolism, while *Coch* and *Prelp* are associated with cell shape and extracellular matrix constitution, respectively, are overexpressed in P60. As for the genes overexpressed in P3, the gene *Mfap2* is component of microfibrils, which are involved in the regulation of growth factors, and several collagen family genes known for their importance in strengthening and supporting tissue structures.

3.3.3 Endothelial Cells

As for the endothelial cells a total of 10 genes were found in both techniques analysis, and all of these are overexpressed in P3 (Table 5) suggesting an higher activity of this particular cell type in the developmental stage rather than in adult.

Table 5 | *Comparison between the common genes found in the analysis of Bulk RNA-seq and endothelial cells in single-cell RNA-seq. All the genes represented have P-adjusted value under 0.05.*

Genes	scRNA-seq (Log2(FC))	Bulk RNA-seq (Log2(FC))
Hmgb2	2.181	3.493
Marcksl1	1.763	4.290
Stmn1	1.664	2.105
Gpihbp1	1.646	2.648
H19	1.493	7.584
Fbln2	1.254	2.991
Ppic	1.246	2.025
Cldn5	1.198	2.503
Apold1	1.102	2.872
Sox11	1.081	4.103

Most of the genes represented have a role in the extracellular matrix and cell adhesion such as, *Cldn5*, *Stmn1* and *Fbln2*, however, *Gpihbp1* is a specific endothelial gene associated with the mediation of lipoproteins transport. Furthermore, the gene *Sox11* despite being the lowest differential overexpressed in P3 is a very known TF that alongside *Sox4* and *Sox12* are responsible for cell survival in developmental tissues, contributing to organogenesis.

3.3.4 Immune Cells

As for the immune cells only contained two relevant genes, one overexpressed in P3 (*Marcksl1*) and the other overexpressed in P60, *Cd74*, which is involved in the stimulation of T-cells.

3.3.5 Ciliogenesis

Despite the DEA not being conducted in the ciliogenesis subcellular type, the canonical markers used for its identification were also searched in the P1vsP60 comparison in the bulk RNA-seq, since they are expected to be overexpressed in P1. And, successfully, the four genes used for the identification of ciliogenesis (*Deup1*, *Shisa8*, *Ccno*, *Mcidas*) in scRNA-seq were overexpressed in the P1 sample with extremely high log2(FC) (4.5, 4.3, 4.5, 6.9, respectively).

3.4 Comparison with other studies

A recent study identified a similar cell type composition in embryonic, adult and aged CP [61], recognizing the epithelial, mesenchymal, endothelial, immune, neuronal and glial cell clusters, although in this particular work the last two were merged. In this same study, the ciliogenesis marker *Shisa8* was also being expressed in multi-ciliated epithelial cells in the embryonic stage decreased in the adult stage, corroborating the results of this work. Furthermore, the epithelial cell type was also the largest cell class in embryo, even though in this project, it was the largest across all stages. At last, several different populations of immune cells were also found, suggesting the diversity of this cell group in the CP, including the subsets B cells, lymphocytes, macrophages, dendritic cells, among others. Although the immune cell cluster was not deeply investigated in this project, several samples demonstrated several clusters associated with this cell type, indicating that the same or other subsets can also be found.

Other studies have also shown that the transition in epithelial cells towards the mature state is associated with an increase in mitochondria's which can also explain the GSEA results from the bulk analysis, since both P10 and P60 displayed enriched energy metabolism gene sets when compared to P1 [71].

The overexpression of the *Prlr* gene in mature epithelial cells has also been investigated, since the CP holds one the main entry point for the hormone prolactin (PRL) through its specific receptors [72]. In this study, the prolactin receptor was overexpressed in the later stages both in the Bulk and single-cell RNA-seq analysis being one of the key genes to differ the developmental from adult stages.

4. Conclusion

From this work, we can conclude that bulk RNA-seq is a viable technique for an initial approach to multicellular tissues, such as the CP, allowing a general identification of its transcriptome across multiple timepoints. The DEA performed allowed the identification of several crucial genes for early development in the brain, as well as in an adult CP. As for the GSEA analysis, it allowed the identification of regular processes and main metabolic pathways associated with each developmental stage. In the early stages, there is a tendency for cell division, adhesion and extracellular matrix organization, while on later stages there is propensity for energy and lipid metabolism. The developmental stage P10 also demonstrated to be a transitional state between adult and embryo since it displayed a wider range of expressed genes.

One other hand scRNA-seq also confirmed the results from the previous technique, providing insights about the cell populations that are responsible for the displayed phenotypes. Furthermore, scRNA-seq also displays a more visual approach to the data which allows a more easing biological perspective to the collected datasets. The possibility of identifying each main cell type in the CP (epithelial, mesenchymal, endothelial, immune and neurons) can provide important information for further studies of the tissue. Ciliogenesis is also an important event that can define observations of early stages of development in epithelial cells. Overall, scRNA-seq provides a more accurate and visual display of results. The conjugation of both techniques offers a more strong and solid perspective about the tissue analysis and should almost always if possible be performed in parallel.

Further studies still need to be performed in order to analyse subpopulations of cell types and potentially unveil more markers and events associated with the development of the CP.

Supplementary Material

All the supplementary material related to this work can be found in the following link:

https://github.com/MiguelMPacheco/Bioinformatics-Thesis-Anexos-2021

Additionally, in this link can also be found the scripts associated with the Bulk RNA-seq analysis, the scRNA-seq analysis of one sample (identical to the others) and their merge/ integration and DEA, and finally, an excel containing information for the comparison of the two techniques (section I).

In order to gain access to the Supplementar Material an email should be sent to:

Miguelpacheco.rt@gmail.com

References

[1] M. A. Lopes Pinheiro *et al.*, "Immune cell trafficking across the barriers of the central nervous system in multiple sclerosis and stroke," *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1862, no. 3, pp. 461–471, 2016, doi: 10.1016/j.bbadis.2015.10.018.

[2] R. Spector, R. F. Keep, S. Robert Snodgrass, Q. R. Smith, and C. E. Johanson, "A balanced view of choroid plexus structure and function: Focus on adult humans," *Exp. Neurol.*, vol. 267, pp. 78–86, 2015, doi: 10.1016/j.expneurol.2015.02.032.

[3] C. Larochelle, J. I. Alvarez, and A. Prat, "How do immune cells overcome the blood-brain barrier in multiple sclerosis?," *FEBS Lett.*, vol. 585, no. 23, pp. 3770–3780, 2011, doi: 10.1016/j.febslet.2011.04.066.

[4] J. V. MD, "Brain Ventricles Location, Role, and Potential Issues." 2020.

[5] M. B. BSc, "Choroid plexus." 2020.

[6] T. Speake, C. Whitwell, H. Kajita, A. Majid, and P. D. Brown, "Mechanisms of CSF secretion by the choroid plexus," *Microsc. Res. Tech.*, vol. 52, no. 1, pp. 49–59, 2001, doi: 10.1002/1097-0029(20010101)52:1<49::AID-JEMT7>3.0.CO;2-C.

[7] V. Martirosian, A. Julian, and J. Neman, *The Role of the Choroid Plexus in the Pathogenesis of Multiple Sclerosis*. Elsevier Inc., 2016.

[8] M. K. Lehtinen, C. S. Bjornsson, S. M. Dymecki, R. J. Gilbertson, D. M. Holtzman, and E. S. Monuki, "The choroid plexus and cerebrospinal fluid: Emerging roles in development, disease, and therapy," *J. Neurosci.*, vol. 33, no. 45, pp. 17553–17559, 2013, doi: 10.1523/JNEUROSCI.3258-13.2013.

[9] R. Di Terlizzi and S. Platt, "The function, composition and analysis of cerebrospinal fluid in companion animals: Part I - Function and composition," *Vet. J.*, vol. 172, no. 3, pp. 422–431, 2006, doi: 10.1016/j.tvjl.2005.07.021.

[10] N. Strazielle and J. F. Ghersi-Egea, "Physiology of blood-brain interfaces in relation to brain disposition of small compounds and macromolecules," *Mol. Pharm.*, vol. 10, no. 5, pp. 1473–1491, 2013, doi: 10.1021/mp300518e.

[11] H. H. Damkier, P. D. Brown, and J. Praetorius, "Cerebrospinal fluid secretion by the choroid plexus," *Physiol. Rev.*, vol. 93, no. 4, pp. 1847–1892, 2013, doi: 10.1152/physrev.00004.2013.

[12] C. Nilsson, M. Lindvall-Axelsson, and C. Owman, "Neuroendocrine regulatory mechanisms in the choroid plexus-cerebrospinal fluid system," *Brain Res. Rev.*, vol. 17, no. 2, pp. 109–138, 1992, doi: 10.1016/0165-0173(92)90011-A.

[13] D. Orešković, M. Radoš, and M. Klarica, "Role of choroid plexus in cerebrospinal fluid hydrodynamics," *Neuroscience*, vol. 354, no. April, pp. 69–87, 2017, doi: 10.1016/j.neuroscience.2017.04.025.

[14] N. L. Hunter and S. M. Dymecki, "Molecularly and temporally separable lineages form the hindbrain roof plate and contribute differentially to the choroid plexus," *Development*, vol. 134, no. 19, pp. 3449–3460, 2007, doi: 10.1242/dev.003095.

[15] K. M. Dziegielewska, J. Ek, M. D. Habgood, and N. R. Saunders, “Development of the choroid plexus,” *Microsc. Res. Tech.*, vol. 52, no. 1, pp. 5–20, 2001, doi: 10.1002/1097-0029(20010101)52:1<5::aid-jemt3>3.3.co;2-a.

[16] C. M. Nielsen and S. M. Dymecki, “Sonic hedgehog is required for vascular outgrowth in the hindbrain choroid plexus,” *Dev. Biol.*, vol. 340, no. 2, pp. 430–437, Apr. 2010, doi: 10.1016/j.ydbio.2010.01.032.

[17] M. K. Lehtinen *et al.*, “The Cerebrospinal Fluid Provides a Proliferative Niche for Neural Progenitor Cells,” *Neuron*, vol. 69, no. 5, pp. 893–905, Mar. 2011, doi: 10.1016/j.neuron.2011.01.023.

[18] V. Silva-Vargas, A. R. Maldonado-Soto, D. Mizrak, P. Codega, and F. Doetsch, “Age-Dependent Niche Signals from the Choroid Plexus Regulate Adult Neural Stem Cells,” *Cell Stem Cell*, vol. 19, no. 5, pp. 643–652, 2016, doi: 10.1016/j.stem.2016.06.013.

[19] E. H. Wilson, W. Weninger, and C. A. Hunter, “Trafficking of immune cells in the central nervous system,” *J. Clin. Invest.*, vol. 120, no. 5, pp. 1368–1379, May 2010, doi: 10.1172/JCI41911.

[20] J. Szmydynger-Chodobska, N. Strazielle, B. J. Zink, J. F. Ghersi-Egea, and A. Chodobski, “The role of the choroid plexus in neutrophil invasion after traumatic brain injury,” *J. Cereb. Blood Flow Metab.*, vol. 29, no. 9, pp. 1503–1516, 2009, doi: 10.1038/jcbfm.2009.71.

[21] J. Szmydynger-Chodobska *et al.*, “Posttraumatic invasion of monocytes across the blood-cerebrospinal fluid barrier,” *J. Cereb. Blood Flow Metab.*, vol. 32, no. 1, pp. 93–104, 2012, doi: 10.1038/jcbfm.2011.111.

[22] D. A. Brown and P. E. Sawchenko, “Time course and distribution of inflammatory and neurodegenerative events suggest structural bases for the pathogenesis of experimental autoimmune encephalomyelitis,” *J. Comp. Neurol.*, vol. 502, no. 2, pp. 236–260, May 2007, doi: 10.1002/cne.21307.

[23] B. Engelhardt and R. M. Ransohoff, “The ins and outs of T-lymphocyte trafficking to the CNS: Anatomical sites and molecular mechanisms,” *Trends Immunol.*, vol. 26, no. 9, pp. 485–495, 2005, doi: 10.1016/j.it.2005.07.004.

[24] M. Sathyanesan *et al.*, “A molecular characterization of the choroid plexus and stress-induced gene regulation,” *Transl. Psychiatry*, vol. 2, no. 7, pp. e139-9, 2012, doi: 10.1038/tp.2012.64.

[25] K. Baruch and M. Schwartz, “CNS-specific T cells shape brain function via the choroid plexus,” *Brain. Behav. Immun.*, vol. 34, pp. 11–16, 2013, doi: 10.1016/j.bbi.2013.04.002.

[26] C. Kaur, G. Rathnasamy, and E. A. Ling, “The choroid plexus in healthy and diseased brain,” *J. Neuropathol. Exp. Neurol.*, vol. 75, no. 3, pp. 198–213, 2016, doi: 10.1093/jnen/nlv030.

[27] P. R. Lowenstein, “Gene Transfer Into the Brain : an Evolutionary and Developmental Perspective,” *Trends Immunol.*, vol. 23, no. 1, pp. 15–17, 2002.

[28] J. M. Keil, A. Qalieh, and K. Y. Kwan, “Brain transcriptome databases: A user’s guide,” *J. Neurosci.*, vol. 38, no. 10, pp. 2399–2412, 2018, doi: 10.1523/JNEUROSCI.1930-17.2018.

[29] K. Van den Berge *et al.*, “RNA Sequencing Data: Hitchhiker’s Guide to Expression Analysis,” *Annu. Rev. Biomed. Data Sci.*, vol. 2, no. 1, pp. 139–173, 2019, doi: 10.1146/annurev-biodatasci-072018-021255.

[30] A. F. Palazzo and E. S. Lee, “Non-coding RNA: What is functional and what is junk?,” *Front. Genet.*, vol. 5, no. JAN, pp. 1–11, 2015, doi: 10.3389/fgene.2015.00002.

[31] S. M. Jazayeri, “RNA-Seq:Advantages, Disadvantages, Problems, Challenges and Applications,” *Conf. Fourth Iran. Conf. Bioinforma.*, 2012.

[32] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, “Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data,” *Elife*, vol. 6, pp. 1–25, 2017, doi: 10.7554/eLife.26476.

[33] R. Salomon *et al.*, “Droplet-based single cell RNAseq tools: A practical guide,” *Lab Chip*, vol. 19, no. 10, pp. 1706–1727, 2019, doi: 10.1039/c8lc01239c.

[34] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, “The Technology and Biology of Single-Cell RNA Sequencing,” *Mol. Cell*, vol. 58, no. 4, pp. 610–620, 2015, doi: 10.1016/j.molcel.2015.04.005.

[35] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines,” *Exp. Mol. Med.*, vol. 50, no. 8, 2018, doi: 10.1038/s12276-018-0071-8.

[36] W. Paper, “The human cell atlas [October 2018],” 2017.

[37] K. Bach *et al.*, “Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing,” *Nat. Commun.*, vol. 8, no. 1, 2017, doi: 10.1038/s41467-017-02001-5.

[38] A. Olsson *et al.*, “Single-cell analysis of mixed-lineage states leading to a binary cell fate choice,” *Nature*, vol. 537, no. 7622, pp. 698–702, Sep. 2016, doi: 10.1038/nature19348.

[39] C. Kim *et al.*, “Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing,” *Cell*, vol. 173, no. 4, pp. 879-893.e13, May 2018, doi: 10.1016/j.cell.2018.03.041.

[40] M. Pavličev *et al.*, “Single-cell transcriptomics of the human placenta: Inferring the cell communication network of the maternal-fetal interface,” *Genome Res.*, vol. 27, no. 3, pp. 349–361, 2017, doi: 10.1101/gr.207597.116.

[41] A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, “Single-cell RNA-seq: Advances and future challenges,” *Nucleic Acids Res.*, vol. 42, no. 14, pp. 8845–8860, 2014, doi: 10.1093/nar/gku555.

[42] L. Liu *et al.*, “Bioinformatics approaches for deciphering the epitranscriptome: Recent progress and emerging topics,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1587–1604, 2020, doi: 10.1016/j.csbj.2020.06.010.

[43] G. Chen, B. Ning, and T. Shi, “Single-cell RNA-seq technologies and related computational data analysis,” *Front. Genet.*, vol. 10, no. APR, pp. 1–13, 2019, doi: 10.3389/fgene.2019.00317.

[44] A. Sathyanarayanan, S. Manda, M. Poojary, and S. H. Nagaraj, “Exome sequencing data analysis,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 164–175, 2018, doi: 10.1016/B978-0-12-809633-8.20094-0.

[45] A. Dobin *et al.*, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013, doi: 10.1093/bioinformatics/bts635.

[46] Y. Liao, G. K. Smyth, and W. Shi, “FeatureCounts: An efficient general purpose program for

assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014, doi: 10.1093/bioinformatics/btt656.

[47] N. F. Lahens *et al.*, "IVT-seq reveals extreme bias in RNA sequencing," *Genome Biol.*, vol. 15, no. 6, pp. 1–15, 2014, doi: 10.1186/gb-2014-15-6-r86.

[48] K. Okonechnikov, A. Conesa, and F. García-Alcalde, "Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data," *Bioinformatics*, vol. 32, no. 2, pp. 292–294, 2016, doi: 10.1093/bioinformatics/btv566.

[49] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016, doi: 10.1093/bioinformatics/btw354.

[50] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015, doi: 10.1093/bioinformatics/btu638.

[51] R. C. T. (2021), "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/." .

[52] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[53] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.

[54] C. R. Bolen, M. Uduman, and S. H. Kleinstein, "Cell subset prediction for blood genomic studies," *BMC Bioinformatics*, vol. 12, no. Figure 1, pp. 1–10, 2011, doi: 10.1186/1471-2105-12-258.

[55] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573-3587.e29, 2021, doi: 10.1016/j.cell.2021.04.048.

[56] T. Ilicic *et al.*, "Classification of low quality cells from single-cell RNA-seq data," *Genome Biol.*, vol. 17, no. 1, pp. 1–15, 2016, doi: 10.1186/s13059-016-0888-1.

[57] S. Durinck *et al.*, "BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, no. 16, pp. 3439–3440, 2005, doi: 10.1093/bioinformatics/bti525.

[58] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nat. Protoc.*, vol. 4, no. 8, pp. 1184–1191, Aug. 2009, doi: 10.1038/nprot.2009.97.

[59] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors," *Cell Syst.*, vol. 8, no. 4, pp. 329-337.e4, Apr. 2019, doi: 10.1016/j.cels.2019.03.003.

[60] G. Hinton and S. Roweis, "Stochastic neighbor embedding," *Adv. Neural Inf. Process. Syst.*, 2003.

[61] N. Dani *et al.*, "A cellular and spatial map of the choroid plexus across brain ventricles and

ages," *Cell*, vol. 184, no. 11, pp. 3056-3074.e21, 2021, doi: 10.1016/j.cell.2021.04.003.

[62] X. Zhang *et al.*, "CellMarker: A manually curated resource of cell markers in human and mouse," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D721–D728, 2019, doi: 10.1093/nar/gky900.

[63] I. Korsunsky *et al.*, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nat. Methods*, vol. 16, no. 12, pp. 1289–1296, 2019, doi: 10.1038/s41592-019-0619-0.

[64] S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann, "The impact of amplification on differential expression analyses by," *Nat. Publ. Gr.*, no. April, pp. 1–11, 2016, doi: 10.1038/srep25533.

[65] B. Bu, C. Kannicht, and W. Reutter, "Novel Cytosolic Binding Partners of the Neural Cell Adhesion Molecule : Mapping," vol. 2, pp. 6938–6947, 2005.

[66] C. H. Mazucanti *et al.*, "Release of insulin produced by the choroids plexis is regulated by serotonergic signaling," *JCI Insight*, vol. 4, no. 23, 2019, doi: 10.1172/jci.insight.131682.

[67] H. Hagiwara, N. Ohwada, and K. Takata, "Cell Biology of Normal and Abnormal Ciliogenesis in the Ciliated Epithelium," vol. 234, pp. 101–141, 2004.

[68] A. Watanabe *et al.*, "Fabp7 maps to a quantitative trait locus for a schizophrenia endophenotype," *PLoS Biol.*, vol. 5, no. 11, pp. 2469–2483, 2007, doi: 10.1371/journal.pbio.0050297.

[69] J. Harris *et al.*, "Socs2 and Elf5 Mediate Prolactin-Induced Mammary Gland Development," vol. 20, no. 5, pp. 1177–1187, 2010, doi: 10.1210/me.2005-0473.

[70] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, "Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium," vol. 12, no. 5, 2011, doi: 10.1093/bib/bbr042.

[71] M. P. Lun, E. S. Monuki, and M. K. Lehtinen, "Development and functions of the choroid plexus–cerebrospinal fluid system," *Nat. Rev. Neurosci.*, vol. 16, no. 8, pp. 445–457, Aug. 2015, doi: 10.1038/nrn3921.

[72] A. R. Costa-Brito *et al.*, "The Choroid Plexus Is an Alternative Source of Prolactin to the Rat Brain," *Mol. Neurobiol.*, vol. 58, no. 4, pp. 1846–1858, 2021, doi: 10.1007/s12035-020-02267-9.

[73] O. Ashenberg, D. Silverbush, and K. Gosik, "ANALYSIS OF SINGLE CELL RNA-SEQ DATA," 2019.