Profiling ageing-associated gene expression alterations in human astrocytes using single cell transcriptomics

Rita Martins Tereso Borges da Silva
ritamtbsilva@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2021

Abstract

Astrocytes are responsible for maintaining the homeostasis of the central nervous system. With ageing, the cellular functions of astrocytes become dramatically compromised. It is plausible that changes in the phenotype of ageing astrocytes can make the brain more vulnerable to injury and age-related diseases (pathological ageing). However, transcriptomic profiling of bulk human brain tissue with RNA-seq fail to discriminate more subtle activity states of astrocytes. As such, this work used publicly available human brain single-cell RNA-seq data to characterise the gene expression signatures of human astrocytes in physiological ageing. These show a clear increase in astrocytic heterogeneity with age, including a group of astrocytes enriched in ageing hallmarks, such as neuroinflammation, excitotoxicity, loss of neuronal support and synaptic homeostasis functions. The enrichment with age of this group of astrocytes has been further validated with independent datasets. Additionally, this work uncovers molecular targets for functional validation, as well as candidate therapeutic compounds for the reversal of pathological aged astrocytes' phenotypes.

Keywords: Astrocytes, Ageing, Neurodegenerative Diseases, Single-cell RNA sequencing

1. Introduction

Astrocytes are a very heterogeneous cell group from a functional and molecular point of view, being essential for neuronal survival and synapse homeostasis. However, occasionally these cells present pathological behaviours that are not protective of the central nervous system (CNS), namely in response to the neuroinflammation concomitant with age. It is plausible that changes in astrocytes phenotype can make the brain more vulnerable to injury and age-related diseases (pathological ageing). However, transcriptomic profiling technologies currently applied in (post-mortem) human brain tissue (such as RNA-seq) fail to discriminate subtle molecular changes that can reflect different activity states of astrocytes.

Furthermore, there are numerous therapies that momentarily improve the symptoms of patients with several neurodegenerative diseases but whose maximum effectiveness is observed in animal models or cell cultures (where they were idealized), suggesting that these systems are not perfect surrogates for modelling age-related illnesses [1]. With the increase in average life expectancy, it is urgent to more deeply understand the role of astrocytes in normal ageing of the human brain with single cell resolution, in order to better grasp how the dynamics of the ageing human brain leads to predisposition to neurodegenerative diseases.

2. Background

2.1. Physiological and Pathological Ageing in the Human Brain

The broad scientific consensus points to ageing as a set of genetic, biological, and environmental factors, which act together and lead to physiological and cognitive changes, compromising cells in their functions [2].

The effect of physiological ageing in the brain is noticeable mainly through cognitive decline. To reach this visible consequence, the ageing brain, like other organs, manifests some cellular and molecular hallmarks of ageing, namely, loss of dendritic spines [2], mitochondrial dysfunction, dysregulated energy metabolism, compromised DNA repair, stem cell exhaustion [3], loss of stem cells in the hippocampus [2], aberrant neural network activity [4] and inflammation [2].

The brain is still quite resilient to physiological ageing, not compromising the elderly to the point of being completely dependent on others and cognitively inept [5]. The subtle changes that occur and may predispose tissues to age-related diseases (cancer, cardiovascular and neurodegenerative disorders such as Alzheimer's Disease (AD) and Parkinson's Disease (PD) [6]) is called pathological ageing. Such may be underlain by neurobiological differences between subjects with the same chronological age [7].

Although the timeline and causality of events as-

sociated with neurodegenerative diseases are not known, and despite these diseases probably having different causes, there are already several clues that associate them with the main hallmarks of ageing [8]. Some authors suggest that the onset of AD includes, among others, inflammation, DNA damage and mitochondrial dysfunction [8]. There are several theories that try to explain the exact relationship between ageing and neurodegenerative diseases. Specifically, one of them admits a continuum between ageing and the appearance of neurodegenerative diseases, such that all ageing will eventually lead to neurodegeneration [8]. Furthermore, inflammation has also been suggested as the primary cause of pathological ageing [9]. It is increasingly agreed among the scientific community that the key to a better understanding of neurodegenerative diseases, and their possible therapeutic reversal, lies in a deeper study of the physiological mechanisms of ageing.

2.2. Astrocytes

The broadest cell type classification in the CNS assumes two distinct groups of cells: neurons and glial cells (oligodendrocytes, microglia, astrocytes) [10]. For many years astrocytes were thought to play a primarily structural role. However, with the increase in scientific knowledge, it is now known that these cells are essential for the proper functioning of neurons and CNS homeostasis [11]. Among astrocytes' main functions, we can highlight the modulation of neuronal activity, synapse homeostasis (including synaptic activity and plasticity, as well as neurotransmitter clearance), provision of trophic factors and nutrients to neurons, and establishment and maintenance of the blood brain barrier (BBB) [12]. New studies also include in the functions of astrocytes the ability to generate brain rhythms and neuronal network patterns [12].

Different astrocytes can present different functions, with this not being an on-off state and some brain regions presenting specific distribution of astrocytic functions [12]. However, astrocytes can also be momentarily activated by various stimuli and diseases, further increasing their complexity and range of functions [12]. Reactive astrocytes, for example, are characterized by morphological, molecular, and functional remodelling, in response to injury, disease, or infection of the CNS [13], and it is not yet clear whether this reactivity is beneficial or detrimental to the homeostasis of the CNS, as they can shift their phenotype to a pro- or anti-inflammatory one.

It is thought that some functional changes that ageing astrocytes undergo may be increasing the pro-inflammatory phenotype of the brain [2]. Specifically, aged astrocytes are thought to activate the complement system, responsible for regulating inflammation through the release of complement factors C3 and C4B, decreasing the strength of the connection between neurons and astrocytes, potentiating memory loss in older people [2]. Furthermore, this may also be associated with the loss of the capacity to maintain synaptic homeostasis, with excitotoxicity being an important hallmark of brains affected by ageing and/or neurodegenerative diseases [14]. Excitotoxicity is mainly the result of prolonged or exacerbated activation of glutamate receptors, caused by the inability of astrocytes to control the levels of glutamate in the synaptic cleft, resulting in loss of neuronal function and cell death [14]. Furthermore, it is known that aged astrocytes have an increase in ROS release, which is related to the oxidative stress theory of ageing [2, 15]. It is also known that aged astrocytes lose part of their ability to maintain the proper functioning of the BBB [2]. Finally, as they are an extremely heterogeneous cell group, any impairment on the astrocyte function will irrevocably impact the function of other neuronal cells, creating feedback mechanisms that result in dysfunction of the entire CNS [2].

Aged astrocytes have also been associated with several neurodegenerative diseases. Specifically, they have been related to AD, as it shares many of the hallmarks of ageing brain and ageing astrocytes, such as oxidative stress, mitochondrial dysfunction, and inflammation [8]. PD has also recently been associated with ageing astrocytes and their consequent loss of function [2].

Since this cell group is very affected by age and given its complexity, it is plausible that there are subtle transcriptional changes associated with ageing astrocytes that remain unnoticed and make the brain more vulnerable to age-related diseases. However, all the ageing astrocyte transcriptomic studies were carried out in mice and/or humans, using RNA-seq of pools of cells [2]. This means that all these changes will reflect an "average" transcriptomic profile of the astrocytes, not having the sensitivity to identify more subtle differences between the transcriptomes of individual cells [16].

2.3. Single Cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is the current gold standard for profiling the transcriptomes of individual cells and thereby inferring their phenotypes. Being a high-throughput technology, it can profile thousands of cells per experiment, allowing at the same time for the study of a single cell transcriptome in an unbiased manner, not targeting specific genes like microarrays do [17].

Given that bulk RNA-seq experiments measure gene expression levels as averages across thousands of cells, if there is high heterogeneity within the

group of cells to be sequenced, transcriptome individualities are lost [16]. With single-cell RNA-seq, we can study each cell individually, obtaining the distribution of gene expression levels across a population of individual cells. It is widely used for discovering new cell states in heterogeneous samples, such as the tumour micro-environment [18].

Single nucleus RNA sequencing (snRNA-seq) is an important variation of scRNA-seq. The single nucleus protocol was developed based on the scRNA-seq protocol to extend its applicability to tissues that cannot be easily dissociated into a single-cell suspension, such as the human brain (given that neurons are highly connected and very long, being difficult to dissociate entirely [19]), or frozen tissues (given that nuclei are better preserved than the whole cell [20]). At the same time, snRNA-seq minimizes the alteration of gene expression that may be introduced by artificial interactions between cells in suspension [21].

3. Materians and Methods

3.1. Data Availability

In this work I used four frozen human brain tissue snRNA-seq datasets publicly available through National Center for Biotechnology Information (NCBI) data repository Gene Expression Omnibus (GEO) [22], using the following search words: *scRNA-seq, epilepsy, Memory, Alzheimer, Alcohol, COVID-19* and *Huntington*.

From these datasets (GSE153807, GSE141552, GSE159812, GSE160936), all control samples were chosen, that is, samples that in principle have no neurological condition that may counfound the analysis. The total snRNA-seq data was composed of cells from 18 samples, comprising an age range from 7 to 91 years old.

Independent human cortex RNA-seq validation datasets and relevant clinical metadata (phs000424.v8.p2) were retrieved from the Genotype-Tissue Expression (GTEx) project [23]. Furthermore, samples associated with dementia, PD, cerebral vascular accident, and unknown cause of death were removed, which reduced the number of samples in roughly 3% (n=255 to n=245).

Both scRNA-seq and GTEx gene expression profiles are publicly accessible as read count tables.

3.2. Software

Most of the work was performed using the `R` software environment for statistical computing and graphics (v4.1.0), and its publicly available packages. Some of the most used R packages in this work were `ggplot2` (v3.3.9) for data visualization, `SingleCellExperiment` [24](v1.14.1) and `Seurat`[25] (v4.0.4) for single cell data handling, `limma`[26] (v3.46.0) for differential expression analysis, `slingshot` [27](v2.0.0) for trajectory inference, `fsgea`[28] (v1.18.0) for gene set enrichment analysis, and `cTRAP` [29](v1.10.0) for drug repurposing. `CIBERSORTx` [30] was also used, in order to estimate cell-type proportions in GTEx independent samples.

3.3. Dimensionality Reduction

As in scRNA-seq read count matrices each gene is a variable, and there are thousands of genes profiled, dimensionality reduction techniques, such as Principal Component Analysis [31] (PCA) and t-distributed stochastic neighbour embedding [32] (t-SNE), are essential. In this work, I chose to use 50 main components for the PCA, and used the `SCE runPCA` function to project the data in these 50 new dimensions, and store the results in an `SCE` object, along with the data and metadata. For the t-SNE representation, the `runTSNE` function of the `SCE` package was used to perform the algorithm and save the new coordinates for visualisation in the `SCE` object, together with the data and metadata. This function also takes a parameter called `perplexity`, which balances the attention between local and global similarities in data, forming clusters, being this set to 30 (the default).

3.4. scRNA-seq data pre-processing

A general pipeline for single cell data pre-processing was used [33](figure 1):

- **Quality Control / Filtering:** Cells with less than 400 counts and 300 unique genes detected were removed. Furthermore, cells with more than 15% of the total reads belonging to mitochondrial genes were also removed, as they may be indicative of low quality cells, where the remaining mRNA may have been lost due to cell lysis or RNA degradation. Only genes with a minimum expression of 5 read counts in at least 10 cells were kept.

- **Doublets:** Doublets occur when more than one cell is captured in the cell sorting protocol. The identification and removal of doublets in this work was done using the `scDblFinder` package from `R`. 4269 putative doublets were removed.

- **Normalisation:** The `computeSumFactors` function (`scran` package version 1.20.1) was used to implement a deconvolution strategy for normalisation. The default window sizes of around 20 cells for low library sizes, and around 100 for high library sizes were used. log_2 transformation was also used, with the addition of 1 pseudocount.

- **Batch Effect Correction:** `Seurat`'s batch effect (that is, variability introduced by the ex-

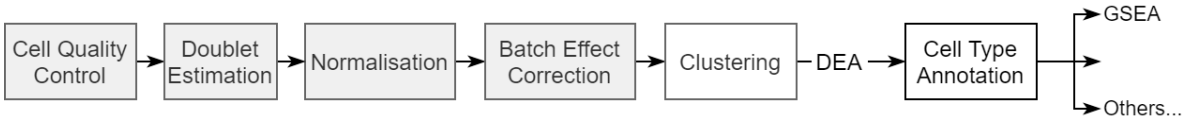

Figure 1: Common workflow for scRNA-seq data processing. Schematic of the processing pipeline used in this work. The pre-processing steps are identified in grey, and were applied to the raw scRNA-seq data count matrix. Adapted from "Analysis of single cell RNA-seq data" (2019) [33].

perimental conditions) correction method [34] was used, with the defaults for each function, including 2000 highly variable genes as the number of features used for finding anchors in the low-dimensional data with the first 30 main components. This correction was only used for visualization and clustering purposes.

3.5. Clustering

One of the most important tasks in analysing single cell data analysis is the definition of clusters of cells (e.g., after t-SNE representation) and the following assignment to cell types. The clustering algorithm was applied using the `Seurat` functions `FindNeighbors` and `FindClusters`, that apply modularity optimization method, or Louvain's method, on top of t-SNE representation of the data. A resolution of 0.4 was chosen, given that the division of clusters with a value higher than 0.4 appears to involve more noise, with the displacements of small cell groups from one cluster to another. This decision was aided by the construction of a cluster tree, using the `clustertree` package.

3.6. Differential Expression Analysis

Differential Gene Expression Analysis (DEA) is essential to infer differences in gene expression between groups of cells, and can be done by linearly model gene expression. I performed differential gene expression analysis (DEA) in multiple ways. First, by modelling gene expression and comparing one cluster against the average of the remaining clusters:

$$GE_x = Cluster_i \times \beta_i \tag{1}$$

Where GE_x is a vector of expression of gene x across cells, and $Cluster_i$ is a logical matrix with an entry of 1 if the cell belongs to cluster i, and 0 otherwise. Given that this matrix has as many columns as clusters and as many rows as cells, and each cell will be in only one cluster, the resulting matrix (design matrix) will be sparse. β_i will be the average expression of gene x in cluster i. A contrast matrix (i.e., a matrix representative of linear combination of the unknown coefficients β_i) was then used to get the differences between a β_i coefficient and the average of the remaining coefficients.

The second way was by comparing two clusters' gene expression. The formulation was equivalent to equation 1, but with further use of a contrast matrix comparing specific pairs of coefficients.

The third way of using the linear models on gene expression in this work was to compare the gene expression profile of each cluster against a baseline one:

$$GE_x = \beta_0 + Cluster_i \times \beta_i \tag{2}$$

Where GE_x is a vector of expression of gene x across cells, β_0 is the expression of the baseline cluster, and $Cluster_i$ is a logical matrix with an entry of 1 if the cell belongs to cluster $i \setminus baseline$, and 0 otherwise. The resulting β_i coefficients will be the log_2FC expression of gene x between each cluster and the baseline cluster.

The limma-voom pipeline was applied to the non-normalized filtered scRNA-seq data (using `edgeR` for normalization) and fit the data to a linear model, using then the moderated t-test (parametric) and empirical Bayes shrinkage of standard errors to assess the statistical significance of the differential expression results. The significance of the results was given by the adjusted p-value for multiple comparisons (Benjamini-Hochberg correction (BH correction)) lower than < 0.05.

3.7. Cell Type Annotation

This task was divided into two main steps: the detailed analysis, where each of the clusters is associated with a cell type, taking into account thresholds based on the percentage of differentially expressed genes between that cluster and the others that are known markers of a cell type [35] ($> 35\%$ of markers of that cell type; $< 8\%$ of each of the remaining cell types; $<60\%$ of unknown markers); and the general analysis, where the smaller clusters are grouped into larger clusters, according to their cell type. `Seurat` v4.0.0's `FindAllMarkers` function finds the markers for each cell cluster against the remaining clusters, using the Wilcoxon Rank Sum Test (non-parametric test). A significance level of adjusted p-value < 0.05 (BH correction) and a magnitude of the difference between clusters of log_2FC above 0.25 were considered, for both the detailed and general analysis.

After these steps, astrocyte cells were selected, and the pipeline was applied again for this subset of 7% of the initial data (going from 209,187 CNS cells (nuclei) to about 13,694 astrocytes).

3.8. Gene Set Enrichment Analysis

The list of differentially expressed genes can be used to perform gene set enrichment analysis (GSEA), that is, compare this ranked list of genes with sets of genes known to be associated with biological pathways and processes. In this work, GSEA was performed using the `fgsea` package [28] to infer phenotypes or biological processes (from GSEA's collection of publicly accessible annotated gene sets) that underlie the biology of each astrocytic cluster. The ranked lists of genes used were the differentially expressed genes in one cluster against the remaining, the differentially expressed genes of one cluster versus the baseline astrocytic cluster, and the loadings of each gene for each principal component.

3.9. Drug Repurposing

Drug repurposing aims to use approved therapeutical compounds for goals different than those they were originally developed to, and surpasses some ethical issues and the expensive time-consuming process of drug development and approval. `cTRAP` [29] can compare an ordered list of differentially expressed genes with known transcriptional alterations caused by gene knockdowns or chemical compounds and find which perturbations are more correlated (positively or negatively) with the phenotype of interest (hereinafter referred to as *phenotype strategy*). Similarly, this tool can, from online databases of drug sensitivity, infer which drugs are the most likely to target cells expressing the marker genes of the cluster with a phenotype of interest (hereinafter referred to as *top gene strategy*).

3.10. Cell-type Deconvolution

Cell-type deconvolution (or digital cytometry) is a technique that allows estimating the proportions of different cell types in bulk samples. This approach is particularly useful in this work, since single-cell protocols may be biased in terms of the proportion of different cell types that are captured, and thus the proportions obtained from individual cell populations may not reflect the true composition of the human brain tissue.

In this work, `CIBERSORTx` [36] was used to perform cell-type deconvolution. Considering data storage limitations in `CIBERSORTx`'s web platform,I removed undefined cell clusters, neuronal clusters whose definition was dubious, and performed random sub-sampling of neurons, oligodendrocytes, and microglia, to remove 4000, 2000 and 1000 cells, respectively. Genes expressing less than 5 counts in at least 100 cells were further removed. For the construction of the cell type signature matrix, an average gene expression threshold (minimum expression) of 0 logFC was chosen. Finally, two groups of astrocytes discovered in this work were removed from the scRNA-seq signature data (type 1 and 6), since it is suspected that they are poor-quality cells. Bulk RNA-seq data were taken from the GTEx project, in this work referred to as "validation data".

4. Results

4.1. Stress as the likely source of variance in ageing astrocytes

It is known that astrocytes are a very heterogeneous group of cells in terms of function and molecular identity. Each of the seven clusters in figure 2 (A) and (B) can be associated with a different type of astrocytes (type 0 to type 6), whose distinctiveness is explored in this section.

It was possible to define cluster 0 as a group of "baseline" astrocytes, that is, whose functions are in accordance with what is expected from a healthy astrocyte (hereafter referred to as "normal" functions). This cluster has an up-regulation of biological processes such as synapse organization, synaptic signalling and neuron development when compared with the other clusters (figure 2 (D), panel "0"), suggesting that the others may have undergone some decline in those processes. Although cluster 0 is the most populated cluster (figure 2 (A)), it has a decrease in proportion in older samples (figure 2 (C)). This reinforces not only that these may be "baseline" astrocytes, present in all samples, but also allows to hypothesise that older samples may be down-regulated in some of this neuronal and synaptic support functions, in accordance with what is already known regarding aged brains.

The main data variance axes appear to be associated with a progression between clusters (figure 2 (B)), given by `slingshot` [27]. Namely, the trajectory parallel to PC1 seems to be associated with clusters 5, 3 and 4. Similarly, the trajectory parallel to PC2 seems to be mainly associated with cluster 2. This suggests two orthogonal axes of some biological response, that drives the main variance in the data, to be further explored.

Both clusters 2 and 5 show a clear enrichment of markers associated with endoplasmic reticulum stress (figure 2 (D)). Namely, they present an up-regulation in the biological processes of granule assembly stress and unfolded protein response, and in the unfolded protein response hallmark, related to ageing. In addition to cluster 2 being enriched in older samples (figure 2 (C)), we also find reactive oxygen species pathways and TGF-β signalling down-regulated in cluster 5 when compared with cluster 2 (figure 2 (D)). These pathways are known to be associated to compensatory immunosuppression after chronic stress, suggesting that cluster 2 expresses chronic stress markers [37, 38], consistent with the oxidative damage theory of ageing [15]. On the other hand, cluster 5 is not predominantly associated with young or old samples. Furthermore, this

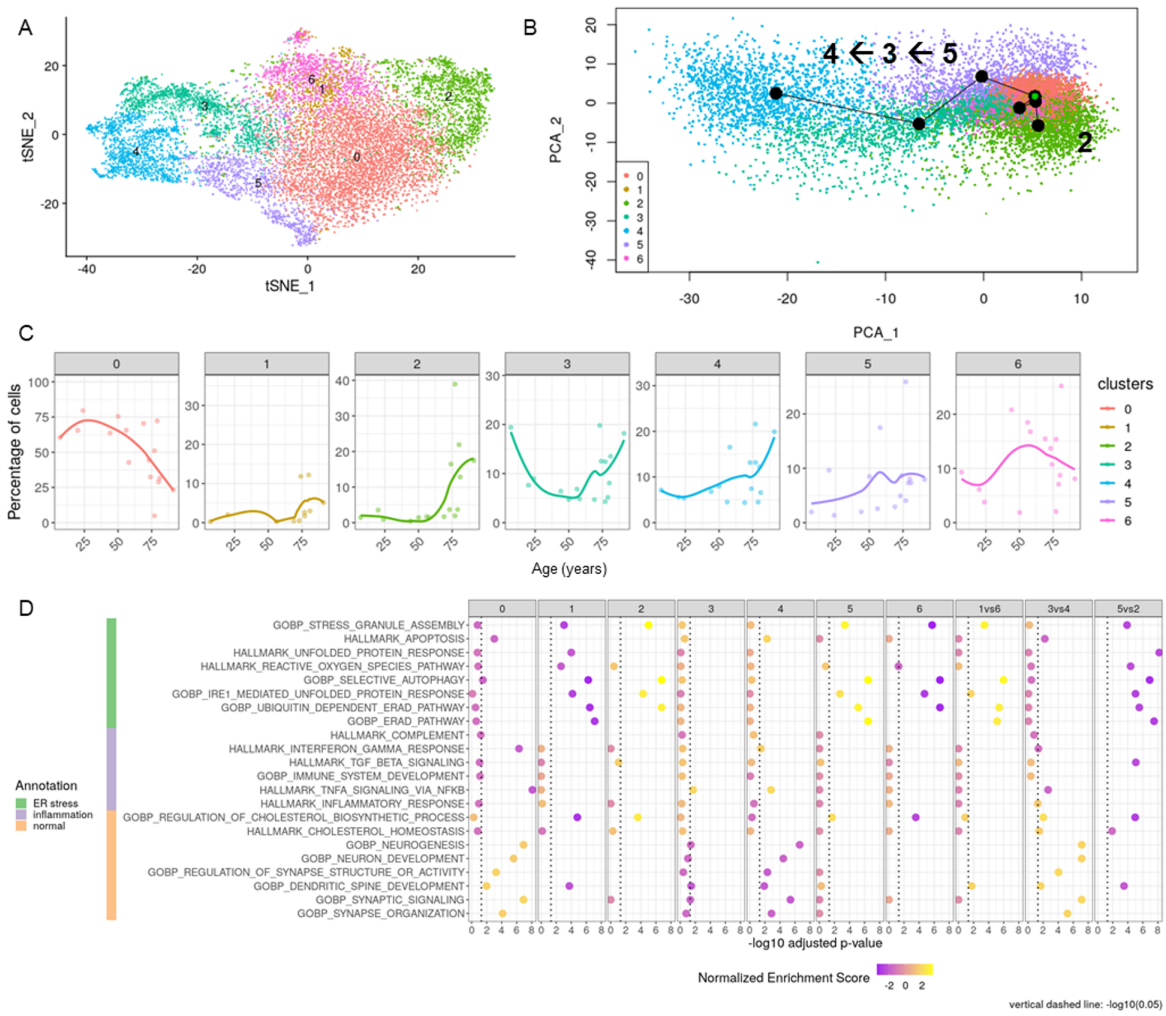

Figure 2: Astrocytic data with clusters 0 to 6. Representation of the astrocyte clusters that make the basis of this work in **(A)** a t-SNE plot and in **(B)** a PCA plot with the first two main components, and with the results of pseudotime inference with potential trajectories between clusters, taking cluster 0 as the "origin" (identified by a green dot); **(C)** Proportions of cells of each cluster along age, with each dot representing a percentage of cells of each cluster in each sample; **(D)** GSEA of pathways, hallmarks and biological processes, associated with ER stress, inflammation and normal astrocytic functions in clusters 0 to 6, as well as specific cluster contrasts. Cluster 0 results have been obtained through DEA of one cluster against the remaining, Clusters 1 to 6 through DEA of each cluster against cluster 0, and the specific contrasts obtained through DEA of one cluster against other.

cluster is associated with the progression of clusters $5 \rightarrow 3 \rightarrow 4$ in the main axis of variance. This may suggest that cluster 5 is an acute ER stress cluster that is somehow related to clusters 3 and 4.

Clusters 3 and 4 appear to be the most similar in gene expression. In functional terms, cluster 4 has enriched neuroinflammation markers and down-regulated markers associated with neuronal support functions and synaptic homeostasis (figure 2 (D)). Furthermore, cluster 3 appears to be more enriched in normal astrocytic functions when compared to cluster 4. Given that these clusters are, together with cluster 5 (acute stress), discriminated along PC1, this might suggest that both clusters 3 and 4 are also acute stress responders characterised by down-regulation of normal astrocytic functions, with 3 being a milder version of 4.

Clusters 1 and 6 appear to have ER stress markers down-regulated (figure 2 (D)), when comparing to the baseline cluster. However, I could not find any other functional information and they are not associated with any meaningful axis of variance in gene expression. As these clusters are are not associated with age, I chose to not further study them in detail.

In short, the enrichment of clusters 2, 3 and 4

with ageing, and their progression in the main axes of variance, may suggest that aged astrocytes show a poor/impaired response to acute stress (clusters 3 and 4), with several associated age hallmarks (inflammatory phenotype) and loss of function, or chronic stress (cluster 2).

4.2. Cluster 4 as a possible target for reversing loss of function in ageing astrocytes

As clusters 2 and 4 are the extremes of the variance and associated with age, it will therefore be in the interest of this work to further study them, as they can be a potential factor for the deregulation of the normal functions of the CNS and predisposition to neurodegenerative diseases.

Combined with the GSEA results suggesting that cluster 4 may be associated with a loss of function by ageing astrocytes in response to acute stress, looking at individual differentially expressed genes therein could give some more specific functional insights into this cluster (figure 3 (A)). Astrocytes in cluster 4 have a deficiency in *SLC1A2* (important for synapse clearance and to prevent excitotoxicity) and *CADM1* (important to maintain functional excitatory synapses). Also, this cluster has an upregulation of *SLC38A1*, a gene encoding for the precursor of GABA and glutamate neurotransmitters. Additionally, this cluster has up-regulated *DCLK1* (axon growth and migration), *DPP10* (synapse homeostasis, binds to voltage-gated potassium channels), *KAZN* (cytoskeletal organization), and *CD44* and TNC (neuroinflammation). Finally, this cluster has down-regulated *NRXN1* (required for efficient neurotransmission), *GPC5* (control of cell division and growth regulation - AD samples have shown to be down-regulated in *GPC5* and *NRXN1* [39]), *CACNB2* (voltage dependent calcium channel protein) and *CABLES1* (important for cell cycle progression, knockdown leads to increased numbers of apoptotic cells). All of the above suggest that cluster 4 of astrocytes exhibits several characteristics known to be associated to pathological ageing (excitotoxicity, downregulation of specific genes, neuroinflammation, etc.).

Cluster 2 does not show enriched astrocyte-related processes in GSEA. Using PC2's ordered list of genes by weights as input to GSEA was used in an attempt to discover more insights into the functions of astrocytes in cluster 2. However, such results were not enlightening, and combined with the fact that PC2 is not exclusively associated with cluster 2, the functional characterization of cluster 2 was not possible.

Although clusters 2 and 4 are both at the extremes of the variance and associated with age, my analyses suggest it is more promising to focus on cluster 4 for subsequent validation and therapeutic exploration. Cluster 4 appears to have a stronger association with age, is at the end of the largest data variance axis and has a more coherent biological gene expression signal (unlike cluster 2, whose functional phenotype, in terms of astrocytic functions, could not be determined). Furthermore, type 4 astrocytes appear to have a stronger association with age in an independent human cortex dataset (*validation data*) than cluster 2 (figure 3 (B) and (C)), a very important validation result that surpasses the potential bias in single-cell RNA-seq data in reflecting true cell proportions.

Differentially expressed genes from cluster 4 (figure 3 (A)), being enriched in older samples, and associated with loss of normal astrocytic function (synaptic maturation, neuronal support) and several characteristic of pathological ageing (inflammation, excitotoxicity), are good candidate genes for validation studies and potential phenotypic reversal for therapeutic purposes.

4.3. Candidate compounds for phenotype reversal of cluster 4

The marker genes of cluster 4, obtained through DEA of cluster 4 against the others, ordered by t value, and with p-value < 0.05, were used as input for `cTRAP`. Trifluoperazine, Niclosamide, Foretinib and Olaparib (figure 3 (D)) were identified as candidates for phenotype reversal of cluster 4 while targeting cells that express cluster 4's marker genes, by having the best FDA-approved compounds for both cTRAP approaches for drug repurposing (Spearman rho coefficient: $<$ -0.01 for the phenotype strategy, > 0.05 for the top gene strategy; product rank coefficient[1]: > 60000 for phenotype strategy, < 100 for top gene strategy).

Trifluoperazine is a drug used for the treatment of schizophrenia for over 50 years. However, a study showed that this drug can slow neurodegeneration by enhancing autophagy in response to stress, in a PD context [40]. Furthermore, although being a drug mainly used for the treatment of parasitic infections, some studies have proposed Niclosamide as a way of attenuating pro-inflammatory and migratory phenotypes of microglia and astrocytes in ALS models [41], and also as having a neuroprotective effect [42]. Foretinib is currently in clinical trials for the treatment of cancer, however it has been proposed to prevent axon degeneration, via preservation of the mitochondria, being thus a candidate for many neurological diseases [43]. Finally, Olaparib is a drug used in the treatment of several types of cancer, namely breast cancer or fallopian tube cancer. A study suggested that the administration of Olaparib in a Huntington's disease model promoted neuroprotection and modulation of the

[1]The **rank product** summarises the individual rankings from cTRAP's comparison methods (Spearman, Pearson and GSEA-based scores)[29].

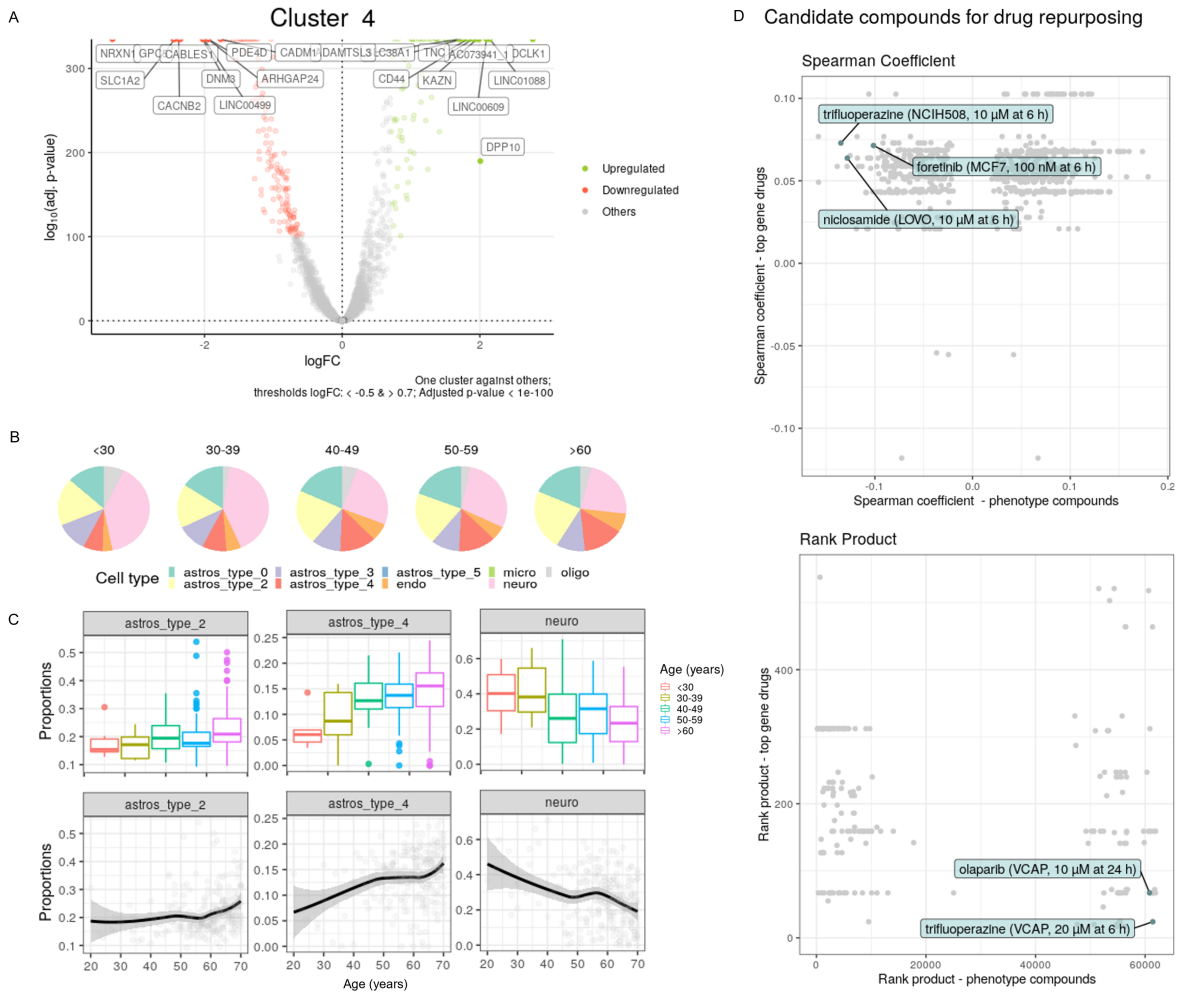

Figure 3: Cluster 4 as a possible target for reversing loss of function in ageing astrocytes. **(A)** Volcano plots of differential expression analysis of astrocytes in cluster 4 against the mean of all other clusters, with some of the most differentially expressed genes highlighted; **(B)** Distribution, by age group, of the proportions of the various cell types, and **(C)** distribution of proportions, by age group and through a general additive model along age (`R geom_smooth` function with default parameters), in neurons and clusters 2 and 4, over the various cortex-independent bulk RNA-seq samples (GTEx); **(D)** Scatter plots comparing the correlation between cluster 4's gene expression changes and those induced by each of CMap's compound perturbations (x axis) and the correlation between the differential gene expression results of cluster 4 and gene expression / drug sensitivity association across all cell lines from *CTRP 2.1* [29] (y axis). The comparisons are performed using Spearman's correlation coefficient and Rank product. Highlighted compounds are FDA-approved candidate reverters of cluster 4's phenotype.

inflammasome activation, resulting in the reduction of neurological deficits and improving the clinical outcomes in neurobehavioural tests [44].

All of these compounds are approved by the FDA and, besides Olaparib, show to be capable of passing the BBB. However there is evidence that these conventional models of the BBB may not predict clinical pharmacokinetics, and thus more studies should be performed on this possibility [45]. Certainly, further validation of these *in silico* results is needed but they are a proof of concept and the basis for future research.

5. Concluding Remarks

Ageing is the strongest risk factor for numerous neurodegenerative diseases, yet the causes that underly the shift from physiological to pathological ageing remains nuclear. Several efforts have been made by the scientific community to discover those causal functional and molecular mechanisms. Astrocytes are a very heterogeneous cell type from a functional and molecular point of view, being essential for neuronal survival and synapse homeostasis. However, occasionally these cells present pathological behaviours that are not protective of the cen-

tral nervous system, namely in response to neuroinflammation concomitant with age. It is plausible that changes in astrocytes' phenotype can make the brain more vulnerable to injury and age-related diseases (pathological ageing). However, technologies currently applied to profile transcriptomes of bulk human brain tissues (such as RNA-seq) fail to detect subtle changes that may allow the identification of different astrocyte activity states. Single-cell RNA-seq allows the study of the transcriptomic profile of each cell individually. Given the complexity of the human brain, the transcriptomic resolution given by this technique allows to identify novel candidate genes and signalling pathways / biological processes characteristic of cells most relevantly contributing to the ageing of brain tissues.

This work culminated in the transcriptomic characterization of a group of astrocytes (type 4) whose enrichment in old samples was identified both in scRNA-seq data and in independent bulk RNA-seq data. Through relationships suggested by a variety of computational tools, this cluster seems to be associated with a down-regulation of neuronal support and synaptic homeostasis functions, in response to acute stress, and also enriched in markers of neuroinflammation and excitotoxicity, being therefore associated with behaviors that are detrimental for the proper functioning of the CNS.

This work scientifically contributed with the discovery of molecular targets for phenotype validation *in vitro* and *in vivo*, as well as candidate therapeutic compounds for the reversal of the pathologically aged astrocytes' phenotype.

5.1. Analysis Limitations and Future Work

First, given that the sample size in this work is relatively small, age may be counfounded with the biological invididuality from each sample. Secondly, the scRNA-seq data used in this work are only from the human cortex. Consequently, there is a lack of regional coverage. Both these caveats could be mitigated if I had access to a greater sample size comprising different areas of the brain.

Another caveat in this analysis was the scarcity of young brain samples. The ideal scenario would be to obtain samples of human brain tissue from healthy young individuals; however, biopsying healthy brains is impossible for obvious ethical reasons.

Although our analyses of gene expression alterations and therapeutic potential of astrocytes have focused mainly on cluster 4, it will still be interesting to further study cluster 3. This cluster seems to be associated not only with PC1 of astrocytic gene expression data but also to PC2. As both of these axes appear to convey different biological responses, the reason for cluster 3 to appear in both could also be interesting to further study in more detail. Furthermore, it would be interesting to expand this study to other CNS cell types. For example, microglia are known to be in close contact with reactive astrocytes [46], and it is possible that microglia also have activation states, some even correlated with cluster 4's enrichment in older samples.

Despite an *in silico* validation of the enrichment of this group of astrocytes with age, functional validation *in vitro* or *in vivo* should also be performed on this group of aged astrocytes. Furthermore, the hypothesis put forward in this work is that these type 4 astrocytes are the result of a response to acute stress (i.e., ageing astrocytes have a greater difficultly adapting to this type of stress), shown by a downregulation of normal astrocytic functions. Due to this, it will be interesting, in addition to emulating their phenotype through genetic editing or under/overexpression of certain genes, to use various stressors (for example pharmaceutical ER stress inducers such as *tunicamycin* and *thapsigargin*, or by physiologically-induced ER stress via glucose deprivation [47]) and to observe the transcriptomic response of astrocytic cell lines. It will also be interesting to study the phenotype associated with the transcriptomic profile of type 4 astrocytes in co-cultures of astrocytes and neurons, to emulate, as far as possible, the characteristic and essential neuronal support environment of astrocytes. 44°C

Acknowledgements

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein was performed at the Disease Transcriptomics (NMorais) Lab of Instituto de Medicina Molecular João Lobo Antunes, Faculty of Medicine of the University of Lisbon (Lisboa, Portugal), during the period February-November 2021, under the supervision of Dr. Nuno Morais. The thesis was co-supervised at Instituto Superior Técnico by Professor Susana Vinga. This work was supported by the GenomePT project [(POCI-01-0145-FEDER-022184), supported by COMPETE 2020 – Operational Programme for Competitiveness and Internationalization (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF)] and the European Molecular Biology Organization [EMBO Installation Grant 3057 to N.B.-M.].

References

[1] I. P. Johnson. Age-related neurodegenerative disease research needs aging models. *Frontiers in Aging Neuroscience*, 7, 2015, doi:10.3389/fnagi.2015.00168.

[2] A. L. Palmer and S. S. Ousman. Astrocytes and aging. *Frontiers in Aging Neuroscience*, 10:337, 2018, doi:10.3389/fnagi.2018.00337.

[3] M. P. Mattson and T. V. Arumugam. Hallmarks of brain aging: Adaptive and pathological modification by metabolic states. *Cell Metabolism*, 27(6):1176–1199, 2018, doi:10.1016/j.cmet.2018.05.011.

[4] J. S. Lee, Y. H. Park, et al. Distinct brain regions in physiological and pathological brain aging. *Frontiers in Aging Neuroscience*, 11:147, 2019, doi:10.3389/fnagi.2019.00147.

[5] D. Murman. The impact of age on cognition. *Seminars in Hearing*, 36(03):111–121, 2015, doi:10.1055/s-0035-1555115.

[6] C. López-Otín, M. A. Blasco, et al. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013, doi:10.1016/j.cell.2013.05.039.

[7] J. H. Cole, R. E. Marioni, et al. Brain age and other bodily 'ages': implications for neuropsychiatry. *Molecular Psychiatry*, 24(2):266–281, 2018, doi:10.1038/s41380-018-0098-1.

[8] Y. Hou, X. Dan, et al. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*, 15(10):565–581, 2019, doi:10.1038/s41582-019-0244-7.

[9] A. A. Simen, K. A. Bordner, et al. Cognitive dysfunction with aging and the role of inflammation. *Therapeutic Advances in Chronic Disease*, 2(3):175–195, 2011, doi:10.1177/2040622311399145.

[10] E. R. Kandel and M. N. Shadlen. *Principles of Neural Science*, chapter 3: Nerve Cells, Neural Circuitry, and Behavior. McGraw Hill, 6 edition, 2021.

[11] O. Gonzalez-Perez, V. Lopez-Virgen, et al. Astrocytes: everything but the glue. *Neuroimmunology and Neuroinflammation*, 2(2):115, 2015, doi:10.4103/2347-8659.153979.

[12] I. Matias, J. Morgado, et al. Astrocyte heterogeneity: Impact to brain aging and disease. *Frontiers in Aging Neuroscience*, 11, 2019, doi:10.3389/fnagi.2019.00059.

[13] C. Escartin, E. Galea, et al. Reactive astrocyte nomenclature, definitions, and future directions. *Nature Neuroscience*, 24(3):312–325, 2021, doi:10.1038/s41593-020-00783-4.

[14] A. Armada-Moreira, J. I. Gomes, et al. Going the extra (synaptic) mile: Excitotoxicity as the road toward neurodegenerative diseases. *Frontiers in Cellular Neuroscience*, 14, 2020, doi:10.3389/fncel.2020.00090.

[15] M. T. Lin and M. F. Beal. The oxidative damage theory of aging. *Clinical Neuroscience Research*, 2(5-6):305–315, 2003, doi:10.1016/s1566-2772(03)00007-0.

[16] F. Chaudhry, J. Isherwood, et al. Single-cell RNA sequencing of the cardiovascular system: New looks for old diseases. *Frontiers in Cardiovascular Medicine*, 6, 2019, doi:10.3389/fcvm.2019.00173.

[17] T. V. Lanz, A.-K. Pröbstel, et al. Single-cell high-throughput technologies in cerebrospinal fluid research and diagnostics. *Frontiers in Immunology*, 10, 2019, doi:10.3389/fimmu.2019.01302.

[18] T. Tammela and J. Sage. Investigating tumor heterogeneity in mouse models. *Annual Review of Cancer Biology*, 4(1):99–119, 2020, doi:10.1146/annurev-cancerbio-030419-033413.

[19] S. R. Krishnaswami, R. V. Grindberg, et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nature Protocols*, 11(3):499–524, 2016, doi:10.1038/nprot.2016.015.

[20] M. Slyper, C. B. M. Porter, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine*, 26(5):792–802, 2020, doi:10.1038/s41591-020-0844-1.

[21] M. R. Alkaslasi, Z. E. Piccus, et al. Single nucleus RNA-sequencing defines unexpected diversity of cholinergic neuron types in the adult mouse spinal cord. *Nature Communications*, 12(1), 2021, doi:10.1038/s41467-021-22691-2.

[22] Gene Expression Omnibus URL `https://www.ncbi.nlm.nih.gov/geo/`. Online; accessed 20 September 2021.

[23] GTEx Portal URL `https://gtexportal.org/home/`. Online; accessed 20 September 2021.

[24] R. Amezquita, A. Lun, et al. Orchestrating single-cell analysis with bioconductor. *Nature Methods*, 17:137–145, 2020, doi:10.1038/s41592-019-0654-x.

[25] Y. Hao, S. Hao, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021, doi:10.1016/j.cell.2021.04.048.

[26] M. E. Ritchie, B. Phipson, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015, doi:10.1093/nar/gkv007.

[27] K. Street, D. Risso, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, page 477, 2018, doi:10.1186/s12864-018-4772-0.

[28] G. Korotkevich, V. Sukhov, et al. Fast gene set enrichment analysis. 2019, doi:10.1101/060012.

[29] B. P. de Almeida, N. Saraiva-Agostinho, et al. *cTRAP: Identification of candidate causal perturbations from differential gene expression data*, 2021. https://nuno-agostinho.github.io/cTRAP, https://github.com/nuno-agostinho/cTRAP.

[30] CIBERSORTx URL `https://cibersortx.stanford.edu/`. Online; accessed 17 September 2021.

[31] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016, doi:10.1098/rsta.2015.0202.

[32] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, pages 2579–2605.

[33] 2 Introduction to single-cell RNA-seq: Analysis of single cell RNA-seq data. Jul 2019. URL `https://scrnaseq-course.cog.sanger.ac.uk/website/introduction-to-single-cell-rna-seq.html`. Online; accessed 22 September 2021.

[34] T. Stuart, A. Butler, et al. Comprehensive integration of single-cell data. *Cell*, 177(7), 2019, doi:10.1016/j.cell.2019.05.031.

[35] A. T. Mckenzie, M. Wang, et al. Brain cell type specific gene expression and co-expression network architectures. *Scientific Reports*, 8(1), 2018, doi:10.1038/s41598-018-27293-5.

[36] A. M. Newman, C. B. Steen, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019, doi:10.1038/s41587-019-0114-2.

[37] S. Fulda, A. M. Gorman, et al. Cellular stress responses: Cell survival and cell death. *International Journal of Cell Biology*, 2010:1–23, 2010, doi:10.1155/2010/214074.

[38] M. K. Brown and N. Naidoo. The endoplasmic reticulum stress response in aging and age-related diseases. *Frontiers in Physiology*, 3, 2012, doi:10.3389/fphys.2012.00263.

[39] P. Preman, M. Alfonso-Triguero, et al. Astrocytes in Alzheimer's disease: Pathological significance and molecular pathways. *Cells*, 10(3):540, 2021, doi:10.3390/cells10030540.

[40] Y. Zhang, D. T. Nguyen, et al. Rescue of pink1 deficiency by stress-dependent activation of autophagy. *Cell Chemical Biology*, 24(4), 2017, doi:10.1016/j.chembiol.2017.03.005.

[41] A. Serrano, S. Apolloni, et al. The S100A4 transcriptional inhibitor niclosamide reduces pro-inflammatory and migratory phenotypes of microglia: Implications for Amyotrophic Lateral Sclerosis. *Cells*, 8(10):1261, 2019, doi:10.3390/cells8101261.

[42] K. C. Bermea, E. A. Casillas, et al. Evidence of a neuroprotective function for niclosamide in human sh-sy5y neuroblastoma and rat PC12 neural cells. *Acta Scientific Neurology*, 3(9):85–94, 2020, doi:10.31080/asne.2020.03.0235.

[43] K. Feinberg, A. Kolaj, et al. A neuroprotective agent that inactivates prodegenerative trka and preserves mitochondria. *Journal of Cell Biology*, 216(11):3655–3675, 2017, doi:10.1083/jcb.201705085.

[44] E. Paldino, V. D'Angelo, et al. Modulation of inflammasome and pyroptosis by olaparib, a PARP-1 inhibitor, in the R6/2 mouse model of huntington's disease. *Cells*, 9(10):2286, 2020, doi:10.3390/cells9102286.

[45] C. Hanna, K. M. Kurian, et al. Pharmacokinetics, safety, and tolerability of olaparib and temozolomide for recurrent glioblastoma: results of the phase i oparatic trial. *Neuro-Oncology*, 22(12):1840–1850, 2020, doi:10.1093/neuonc/noaa104.

[46] S. A. Liddelow, K. A. Guttenplan, et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, 541(7638):481–487, 2017, doi:10.1038/nature21029.

[47] C. M. Oslowski and F. Urano. Measuring ER Stress and the unfolded protein response using mammalian tissue culture system. *The Unfolded Protein Response and Cellular Stress, Part B Methods in Enzymology*, page 71–92, 2011, doi:10.1016/b978-0-12-385114-7.00004-0.