UNIVERSITÀ DI TRENTO

Analysis of RNA-sequencing data taken from microglial cells of *rTg4510 tau* transgenic and wild-type mice

Maurizio Gilioli

1 INTRODUCTION

Microglial cells are tissue macrophages of the central nervous system (CNS). They help in maintaining the natural morphology of the tissue during adulthood, and, throughout the development stage of the organism, they guide the regular tissue morphogenesys, importantly controlling the formation and depletion of synapses[1]. Microglial cells are essential to carry out immune innate responses inside the CNS, but they also provide a series of beneficial effects promoting cell health, cell genesis and dendritic growth[1]. Following the formation of damages inside the CNS, microglial cells are able to assume an activated form, which is more capable of proliferation, movement and other morphological changes. Microglial cells contribute to the normal tissue recovery from lesions and infections; they take action to restore the clearance of the environment[2,3] . However, it was observed in samples isolated from patients affected by neurological degenerative diseases that not always activation of microglial acts beneficially: in those cases the activation is deregulated and consequently causes damage to the nervous tissue. microglial cells are generally considered to be able to assume two possible phenotypes: *M1/pro-inflammatory* and *M2/tissue-repai*r[2,3] . Among the mechanisms generally associated with the onset of neurodegenerative diseases, such as Parkinson and Alzheimer, *tau* accumulation has been shown to be one of the most important. *tau* is a protein related to the cytoskeleton, and its activity is fundamentally dependent on the grade of phosphorylation: misfolded hyperphosphorylated *tau* isn't able to interact correctly with microtubules, and eventually aggregates in the cytoplasm[1]. Despite its association with neurodegenerative diseases, the mechanisms linking the pathological *aggregation* of *tau* with synaptic dysfunction and neurodegeneration are poorly understood[4] . To study tauopathies, mice models have been used in history, despite the fact that some limitations have been discovered when transposing the obtained information to the humans[4] . The current study has been conducted by using *rTg4510* mice, that were transgenically modified to express an human 4-repeat *tau* containing an FTLD-17 associated mutation (P301L), which is responsible for age-dependent pathological *tau* accumulation[5].

1.1 Aim of the project

The project was done as a requirement for the course of **"Network based data analysis"**, held by professor Mario Lauria for the degree of Quantitative and Computational Biology in the university of Trento. The aim of the project was to analyze the expression difference between cells isolated from normal and transgenic tissues. Samples of *all-ages* were initially compared (***all-ages case***), successively, only young samples were tested for differentially expressed genes (***young-samples case***), The results of the analysis are reported in the Results section and in the Supplementary file.

2 METHODS

2.1 Data retrieval and Image plotting

- **Data retrieval:** The analyzed data were collected from the Recount3 Database. The repository code is **SRP172787** (GEO page), and it was uploaded into R thanks to the 'recount3' version 1.6.0 package, publicly available[6] . The data consists of RNA sequencings of microglial cells of forebrain tissues dissected from *rTg4510 tau* transgenic and wild-type famale mice. A total of 96 cells, 12 for each age and genotype combination, were isolated from four age groups of mice (2-, 4-, 6-, and 8-months old). Those cells were analyzed with the aim of capturing the longitudinal gene expression changes corresponding to varying levels of pathology, from minimal *tau* accumulation to massive neuronal loss.
- **Image plotting**: Most of the images were generated thanks to the ggplot2 R package[7].

2.2 Data pre-processing

The raw counts data downloaded from the recount repository were changed through adequate pre-processing steps, performed as follows:

1. Conversion of the base-pair counts (provided by the recount3 database) into read counts.

2. *(First filtering)* Filtering of genes with zero mapped reads for more the 20% of the samples.

3. Normalization following the GEMM method.

4. *(Second filtering)* Filtering of genes with *cpm* values below 0.5 for more than the 20% of the samples.

5. Two consecutive log_2 transformations of the read count data.

The results of this pre-processing steps are shown in section "Data pre-processing" of the Supplementary file. After the first filtering, 30351 out of the total 55421 were retained. After the second one, **22472** were remaining.

2.3 Main study procedure

Two main questions, highlighted in the Aim section 1.1 (comparison between samples of *all-ages*, comparison between *young-samples* (2-4 months)), were asked throughout the project. First of all, labels were introduced and used to divide the samples in distinguished groups ("Transgenic" and "Wild-type"), then, a cycle of analysis was performed for each question, starting with Principal component analysis (**PCA**). Successively, unsupervised learning methods were used, including **K-means clustering** (figures 1b and 4b) and **Hierarchical clustering**. Both the mentioned unsupervised methods were performed by using the complete set of genes (22472 genes). Hclustering results are shown in the section "Hclustering results" of the Supplementary file.
Then, Wilcoxon rank sum tests[8] were done to test each gene. The returned *p*-values were adequately adjusted through the *p.adjust* R method and used to filter the original matrix. The genes retained after this filtering process were 4991 in the *all-ages* case study, 1342 in the *young-samples* case. Exploiting the importance values (Supplementary file, section "Importance values") obtained after a random forest analysis performed over the restricted datasets, the 25 genes with the highest importances were used to generate an Heatmap of the samples (figures 2 and 5).

Using the reduced matrix, seven supervised learning methods were applied, elencated in the following lines:

- cross-validated Random Forest (10 partitions).
- cross-validated LDA (10 partitions).
- repeated cross-validated Random Forest (10 partitions; 10 repetitions).
- repeated cross-validated LDA (10 partitions; 10 repetitions).
- cross-validated Lasso regression (10 partitions).
- bootstrapping Random Forest (10 iterations).
- bootstrapping LDA (10 iterations).

2.3.1 Network generation

the RScudo package and the cytoscape program[9,10] were used in combination to generate explanatory networks. At first, the best values of nBottom and nTop were evaluated. Those were then passed to the scudoTrain function. The resulting ScudoResults object was taken as input by the scudoNetwork function with an N value of 0.35 and the resulting network was manipulated with cytoscape. Results are plotted in figure 7.

2.3.2 Enrichment analysis

Enrichment analysis was performed through g:profile,[11] the enrichGO function[12], and the pathfindR[13] package. The search with pathfindR was done using the *mmu_STRING* protein interaction network (PIN) and the *mmu_KEGG* gene set associated with the library. The enrichment profiling results obtained with g:profiler, enrichGO and pathfindR gave the results shown in the "Enrichments results" section of the Suppl. file.

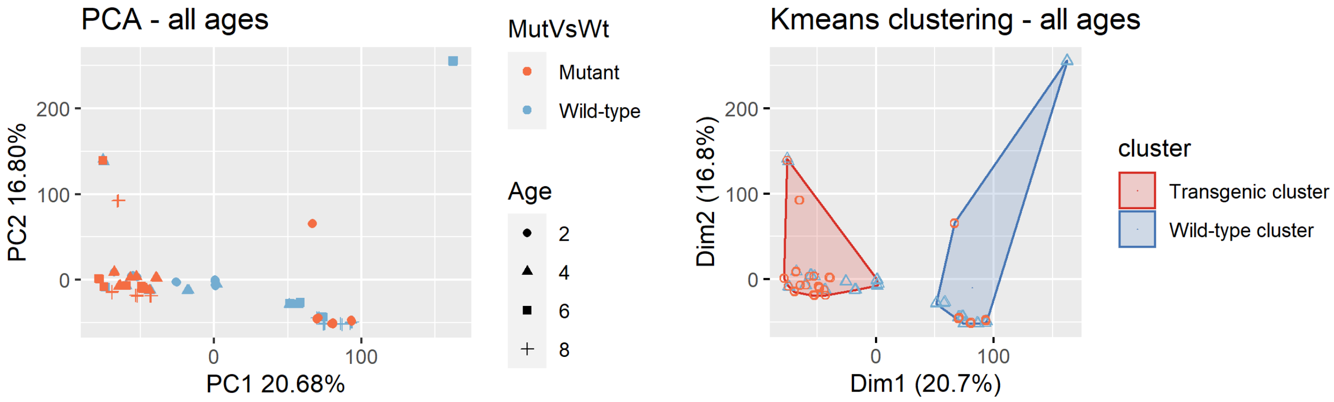
3 RESULTS

3.1 Differential expression analysis - *All-Ages* case

3.1.1 First analysis and unsupervised methods - All-Ages case

The PCA analysis, applied to all the samples, grouped on the base of the genotype, gave good results in terms of genotype separation, as shown in figure 1a. The first two components only explained respectively the 20.68% and the 16.80% of the total variance. Two clusters are visibly identifiable: one on the bottom-right corner, at around $PC1 = 100$ and $PC2 = -50$, and the other at around $PC1 = -50$ and $PC2 = 0$. By distinguishing the dots on the base of the age and of the genotype, it can be observed that most of the transgenic samples of two months of age (orange circles in figure 1a and 1b), were not visibly separable from the cluster mainly containing control samples on the bottom right of the graph, suggesting the presence of a larger difference between older transgenic and wild-type samples.

The presence of a large divergence between transgenic and wild-type samples was confirmed building the hierarchical clustering (Suppl. file) and the heatmap graphs (figure 2): all the samples were associated to two main groups.



(a) PCA analysis, *all-ages* case

(b) Kmeans clustering, *all-ages* case

3.1.2 Filtering process - all-ages

After the filtering process performed with Wilcoxon rank sum tests, 4991 genes were found to have a p-adjusted value lower than 0.05, and therefore those were used in the subsequent analysis.

3.1.3 Supervised learning classification - all-ages

Supervised methods for classification gave surprisingly good results: all of them attested values of accuracy around 1 (100%), reported in figure 3. However, the models made through the LDA method gave generally less accurate results. Nevertheless, also in these cases accuracy was around 0.95%. All together, these results suggest that it is possible to classify correctly a sample coming from a transgenic animal.

3.1.4 Enrichment analysis - all-ages

136 genes were found with interaction in the PIN used by pathfindR out of the 250 given in input. Among the terms listed by pathfindR, the following were found to be significantly enriched: ***"Ribosome"*** (p-val: $2.2e-42$), ***"Coronavirus disease - COVID-19"*** (p-val: $2.1e-40$), ***"Lysosome"*** (p-val: $1.6e-12$), ***"Apoptosis"*** (p-val: 1.6e-05), ***"Toll-like receptor signaling pathway"*** (p-val: $4.8e-04$), ***"Antigen processing and presentation"*** (p-val: $5.0e-04$), ***"TNF signaling pathway"*** (p-val: $7.2e-04$), ***"Oxidative phosphorylation"*** (p-val: 9.1e-4), ***"Parkinson disease"*** (p-val: 8.1e-3), ***"Prion disease"*** (p-val: 8.6e-3).

3.2 Differential expression analysis - *young-samples*

3.2.1 First analysis and unsupervised methods - young-samples case

In figure 4a are shown the results of PCA. As already mentioned before, transgenic samples of two months of age seem to cluster together with those belonging to the control. The heatmap shown in figure 5 highlights a clear distinction through the samples into three groups: one grouping the control samples, one grouping all the samples 2 months old and a last one groping all the 4 months old transgenic samples.

3.2.2 Filtering process - young-samples

After the filtering process performed with Wilcoxon rank sum tests, 1342 genes were found to have a p-adjusted value lower than 0.05, and therefore were used to in the subsequent analysis.

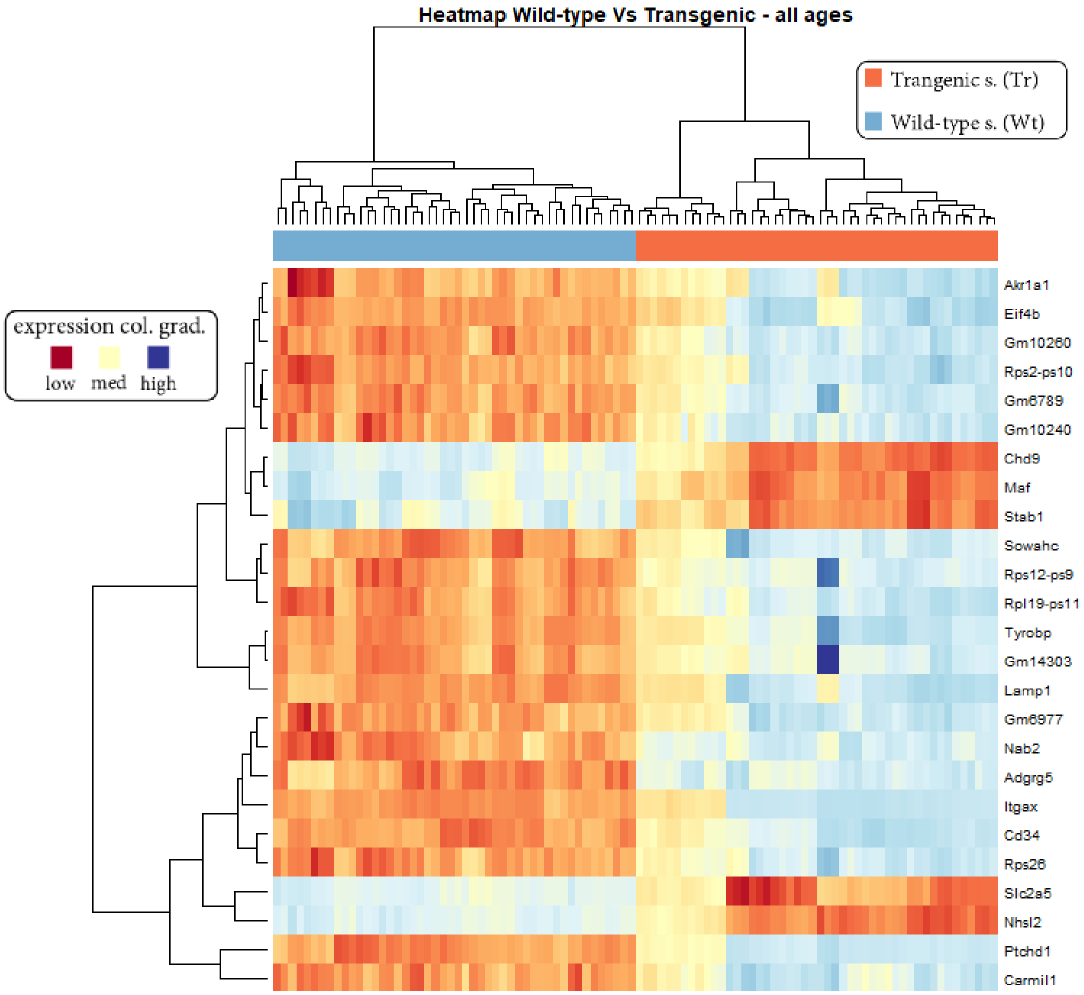
3.2.3 Supervised learning classification - young-samples

The results are shown in figure 6. All the classification methods showed perfect accuracy. These signify that it is possible to precisely classify samples to the correct genotypes.

3.2.4 Network analysis - young-samples

The network obtained through RScudo identified two main clusters, composed by transgenic samples and Wild-type samples respectively. The resulting network is shown in figure 7.

Figure 2. Heatmap, *all-ages* case



3.2.5 Enrichment analysis - young-samples

123 genes were found with interaction in the PIN used by pathfindR out of the 200 genes given in input. Among the terms listed by pathfindR, the following were found to be significantly enriched: ***"Ribosome"*** (p-val: $3.8e-25$), ***"Coronavirus disease - COVID-19"*** (p-val: $3.0e-21$), ***"Lysosome"*** (p-val: $1.2e-07$), ***"Toll-like receptor signaling pathway"*** (p-val: $1.5e-6$) ***"Cytokine-cytokine receptor interaction"*** (p-val: 4.7e-6), ***"Cell adhesion molecules"*** (p-val: $6.4e-6$), ***"Chemokine signaling pathway"*** (1.1e-5), ***"TNF signaling pathway"*** (p-val: 3.0e-04), ***"T cell receptor signaling pathway"*** (p-val: 5.6e-04), ***"Oxidative phosphorylation"*** (p-val: 9.1e-4), ***"Apoptosis"*** (p-val: 1.3e-03), ***"Parkinson disease"*** (p-val: 8.1e-3), ***"Prion disease"*** (p-val: 8.6e-3).

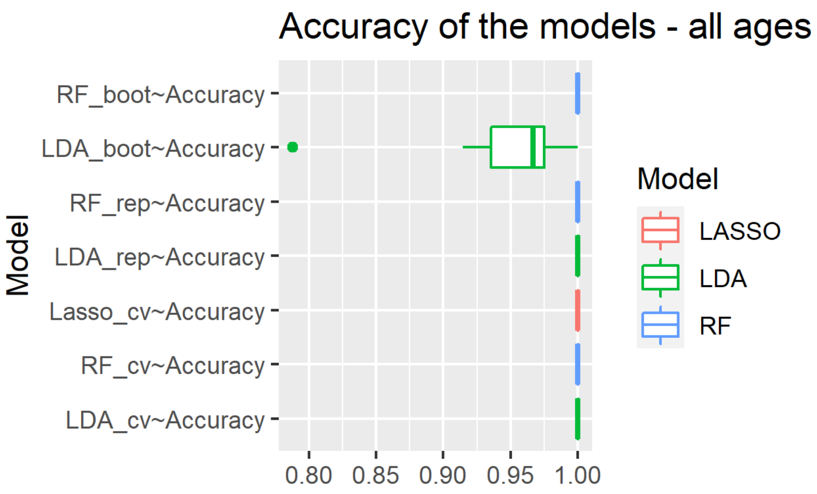
4 Discussion and conclusions

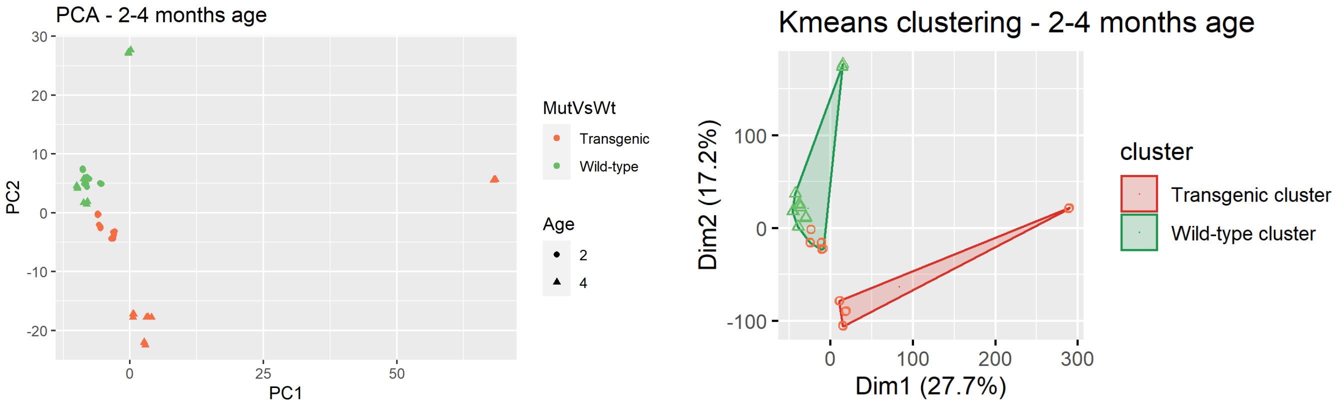
After having pre-processed the data, two biological questions were made. In particular, it was of my interest to compare firstly all the samples of all the ages performing differential gene expression analysis on the basis of the Genotype characteristic. Secondly, I decided to compare only the samples two or four months old, with the objective of finding differences between young and old samples.

As shown in figures 3 and 6, it was possible to generate highly accurate models for performing classifications. All of them gave accuracies ranging around 1, however, LDA generally retrieved lower results.

Performing the gene enrichment analysis, several genes passed as input for pathfindR were not recognised inside the mmu_STRING PIN; this problem could be due to the fact that the mouse relative PINs are generally less studied and characterized than those of humans, and consequently, it is generally harder to make analysis on them. Despite those limitations, strongly enriched terms were found in both the cases. As I expected before starting the project, one of the

Figure 3. Accuracy results - *all-ages*





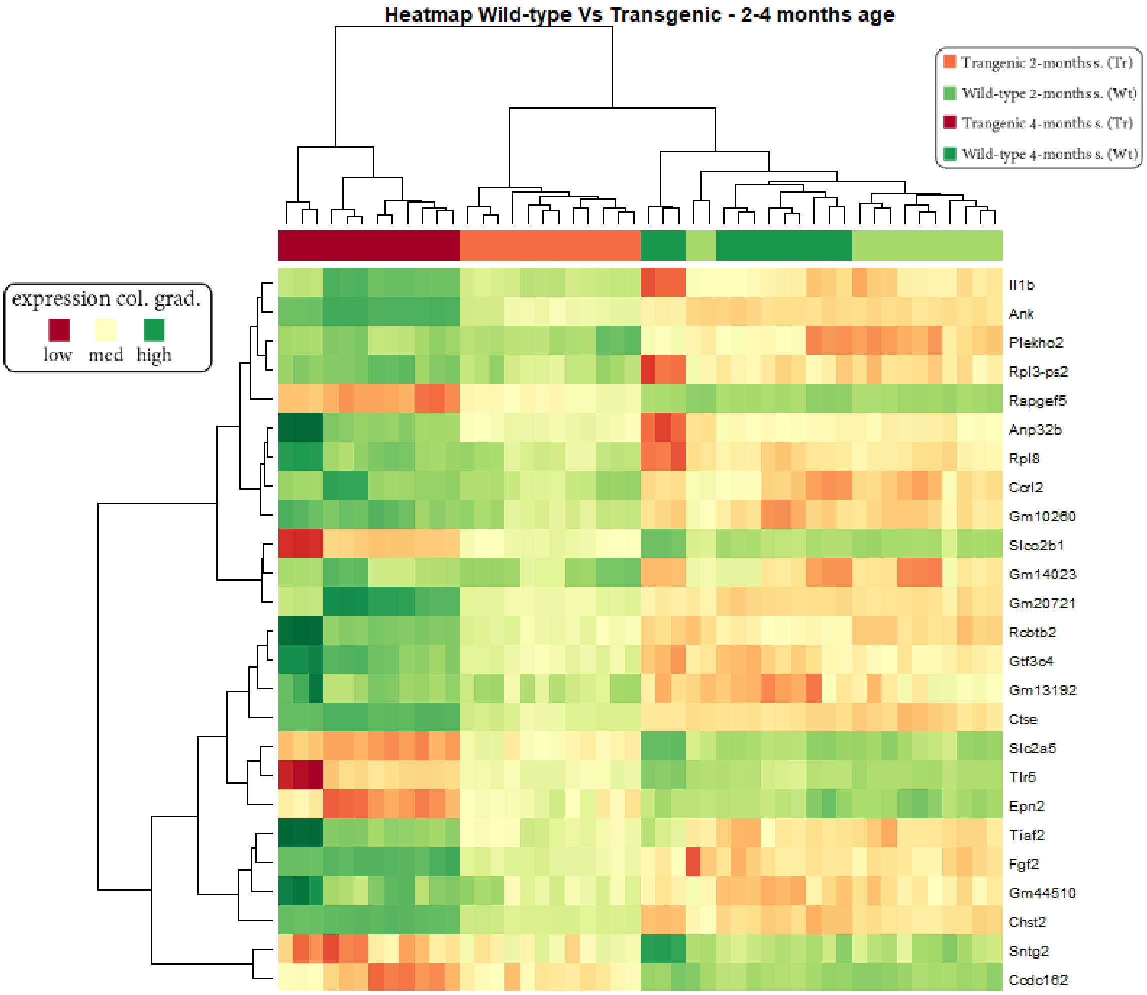
(a) PCA results, *young-samples* case

(b) Kmeans results, *young-samples* case

pathways that seems to be altered mostly is the one regarding the expression of ribosomal proteins: as microglial cells result more activated, it's reasonable to think that an activated state is also associated with an higher cellular activity, usually followed by an higher expression of the ribosomal constituents. Possibly, the enrichment of other paths could be explained in this way, for example the "Oxidative phosphorylation" path, as well as the terms associated with the innate immune system and the production of TNF[2,3]. The "Lysosome" term was found enriched in both the cases too; as *tau*-derived pathologies are generated by protein uncontrolled accumulation, lysosomes are expected to be not properly functioning in these situations[14]. From the analysis I performed, significant enrichment was found for the Parkinson and the Prion neurodegenerative diseases. Parkinson involved genes in the young case were *Gnas*, *Ube2j1*, *Uqcr10*, *Atp5a1*; in the *all-ages* case instead those involved were *Uba52*, *Cox7a2l*, *Cox7b*.
Cox7a2l, *Cox7b*, *Uqcr10* and *Atp5a1* are parts of the electron transport chain[15], while instead *Uba* genes are involved in the ubiquitination process.

Within younger samples, signaling pathways involving cytokines and their receptors are particularly altered. It is a recent finding that chemokines, other than being implied in homeostatic brain functions and developmental processes, are also involved in neurodegeneration and neuroinflammation processes. *Cxcl10*, for example, has been seen to be overexpressed in the context of severe neurological conditions, and in particular, it was found in Alzheimer patients a positive correlation between its concentration and cognitive impairment[16,17]. Microglial cells are retained to be responsible for the production of these chemokines, which induce the recruitment of peripheral blood monocytes and glial cell activation[18]. Importantly, Apoptosis is another term which seems to be differentially enriched in the two cases. Anti-apoptotic proteins pertaining to the *BCL-2* family (*Bcl2a1d*, *Bcl2a1a*, *Bcl2a1b*) are overexpressed, making microglial cells less sensitive to apoptosys; for this reason, microglial cells supernumerary is generally associated with

Figure 5. Heatmap, *young-samples* case



neurological inflammations[3]. Despite some important limitations of this study project, which is limited to female mice and takes into consideration only a specific type of brain cell, it was possible to assess evident differences in terms of gene expression between pathological tissue-derived cells and wild-type cells, explainable through the currently available literature. Finally, and most importantly, it was possible to find differences between different age-stages.

References

1. Španić, E., Langer Horvat, L., Hof, P. R. & Šimić, G. Role of Microglial Cells in Alzheimer's Disease Tau Propagation. *Front. Aging Neurosci.* **11**, 271, DOI: 10.3389/fnagi.2019.00271 (2019).
2. Friedman, B. A. *et al.* Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models. *Cell Reports* **22**, 832–847, DOI: 10.1016/j.celrep.2017.12.066 (2018).
3. Wang, H. *et al.* Genome-wide RNAseq study of the molecular mechanisms underlying microglia activation in response to pathological tau perturbation in the rTg4510 tau transgenic animal model. *Mol. Neurodegener.* **13**, 65, DOI: 10.1186/s13024-018-0296-y (2018).
4. Patel, H. *et al.* Pathological tau and reactive astrogliosis are associated with distinct functional deficits in a mouse model of tauopathy. *Neurobiol. Aging* **109**, 52–63, DOI: 10.1016/j.neurobiolaging.2021.09.006 (2022).
5. rTg(tauP301L)4510 | ALZFORUM. https://www.alzforum.org/research-models/rtgtaup301l4510.
6. Collado-Torres, L. Recount3: Explore and download data from the recount3 project. Bioconductor version: Release (3.15), DOI: 10.18129/B9.bioc.recount3 (2022).
7. Wickham, H. *et al.* Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (2022).
8. CRAN - Package WilcoxCV. https://cran.r-project.org/web/packages/WilcoxCV/index.html.
9. Ciciani, M., Cantore, T., Colasurdo, E. & Lauria, M. rScudo: Signature-based Clustering for Diagnostic Purposes. Bioconductor version: Release (3.15), DOI: 10.18129/B9.bioc.rScudo (2022).

Figure 6. Accuracy results - *young-samples* case

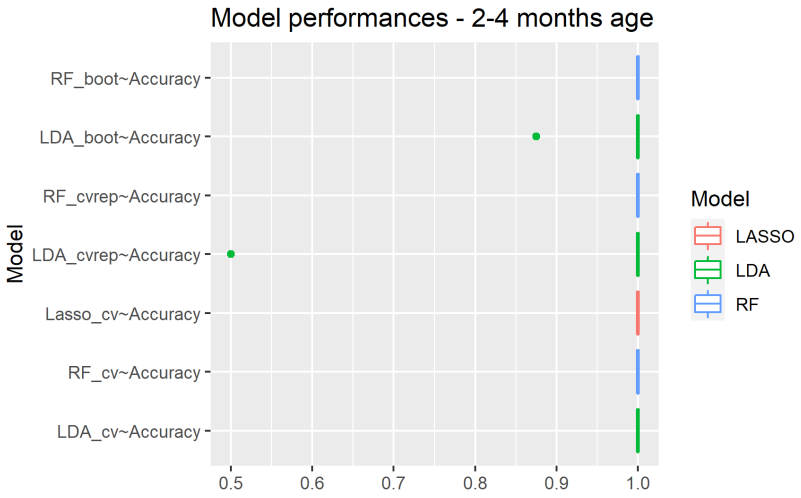
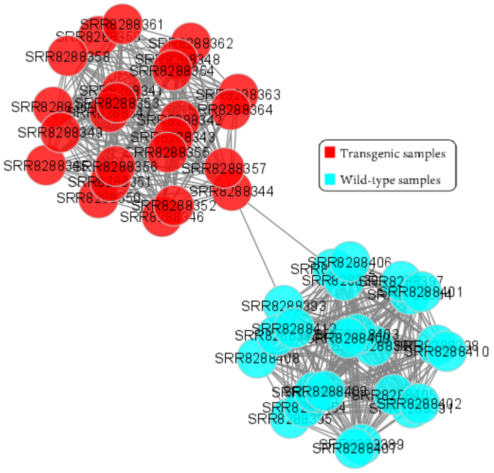


Figure 7. Network of *young-samples* case



10. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504, DOI: 10.1101/gr.1239303 (2003).

11. Raudvere, U. *et al.* G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198, DOI: 10.1093/nar/gkz369 (2019).

12. Bioconductor - clusterProfiler. https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html.

13. Ulgen, E. & Ozisik, O. pathfindR: Enrichment Analysis Utilizing Active Subnetworks (2021).

14. Bourdenx, M. & Dehay, B. What lysosomes actually tell us about Parkinson's disease? *Ageing Res. Rev.* **32**, 140–149, DOI: 10.1016/j.arr.2016.02.008 (2016).

15. Li, J.-L. *et al.* Mitochondrial Function and Parkinson's Disease: From the Perspective of the Electron Transport Chain. *Front. Mol. Neurosci.* **14**, 797833, DOI: 10.3389/fnmol.2021.797833 (2021).

16. Di Castro, M. A. *et al.* The chemokine CXCL16 modulates neurotransmitter release in hippocampal CA1 area. *Sci. Reports* **6**, 34633, DOI: 10.1038/srep34633 (2016).

17. Zuena, A. R., Casolini, P., Lattanzi, R. & Maftei, D. Chemokines in Alzheimer's Disease: New Insights Into Prokineticins, Chemokine-Like Proteins. *Front. Pharmacol.* **10** (2019).

18. Tian, J. *et al.* Specific immune status in Parkinson's disease at different ages of onset. *npj Park. Dis.* **8**, 1–8, DOI: 10.1038/s41531-021-00271-x (2022).