Leveraging single-cell and bulk RNA sequencing data to build microglia aging clocks

Luvna Dhawka and Sneha Jaikumar
COMP 683: Computational Biology

Abstract

Microglia are the brain's primary immune cells, significantly influencing brain development, homeostasis, and injury response[1]. These cells display variable phenotypes and functions as one ages, with the underlying mechanisms responsible for this remaining largely unexplored[2]. Recent studies have linked microglia subtypes to processes involved in neurodevelopment, aging, neurodegeneration, and brain injuries[3–5]. Our long-term goal is to comprehensively characterize the frequency and functional features of these cells across different conditions and leverage this information to build predictive models. Specifically for this project, we will use single-cell RNA sequencing and bulk RNA sequencing data from microglia isolated from mice and humans to identify subtypes and features/genetic markers predictive of age.

Introduction

Problem Motivation

Our project strives to improve the current understanding of the role of microglia in the aging process and mechanisms of age-related diseases, especially in the neuroimmune landscape. We want to explore how this cell contributes to neurodevelopment, aging, neurodegeneration, and brain injury. This knowledge can have a significant impact on various aspects of brain health. For instance:

- Improved insights into the development and progression of neurodegenerative conditions and other age-related diseases could lead to the accelerated development of new diagnosis tools, therapies, and drug discovery and development.
- A microglia-aging clock could help screen individuals for risk of cognitive decline, enabling earlier diagnosis and intervention.
- Successfully implementing the aging clock for microglia and potentially other less-invasively accessible cell types could have implications for personalized medicine. Treatments can be adapted based on an individual's biological age predicted by the clock rather than their chronological age.

Previous work focused on solving this problem/ giving background on this problem

Related works include studies highlighting the exploration of cell-type specific aging clocks, the roles of microglia in aging, neurodevelopment, and neurodegeneration, and the use of single-cell RNA sequencing to identify distinct microglia subtypes. Below are summaries of these works categorized under these headings.

- Cell-type specific aging clocks

Buckley, M. T. et al. Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. Nat Aging 3, 121–137 (2023).

Buckley et al. developed "cell-type-specific aging clocks" using single-cell transcriptomics to quantify aging and rejuvenation in specific brain regions. The main methods involved were:

1. Sc-RNA seq of brain cells from mice across a wide age range to capture cell-type diversity.
2. Training machine learning models (lasso, elastic net) on this data to predict chronological and biological ages of the cells based on gene expression profiles.
3. Validating the aging clocks through cross-cohort validation.
4. Applying the clocks to datasets from mice subjected to interventions like heterochronic parabiosis and exercise to test for rejuvenating effects at the transcriptomic level.
5. Testing the generalizability of the clocks across different brain regions and species.

The study concluded that these aging clocks could effectively quantify biological and chronological ages, as well as the rejuvenating effects of interventions, offering an important tool for understanding aging and evaluating anti-aging strategies at the cellular level.

- Microglia in Neuroimmune Regulation

Harry, G. J. Microglia During Development and Aging. Pharmacol Ther 139, 313–326 (2013).

This paper reviews the roles of microglia, the brain's immune cells, throughout human life.

Key points made are:

1. Microglia regulate brain development, maintenance, injury response, and repair.
2. During development, they aid neurogenesis, pruning, and differentiation.
3. With aging, microglia undergo functional declines contributing to neurodegeneration.
4. Activated microglia can release beneficial or harmful substances.
5. Their dysfunction is implicated in neurodegenerative diseases like Alzheimer's.

The review integrates findings from the literature outlining microglia's essential roles in developmental processes and the aging brain, suggesting potential therapeutic targets for enhancing brain resilience.

- Microglia and neurodegeneration

Gao, C., Jiang, J., Tan, Y. & Chen, S. Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. Sig Transduct Target Ther 8, 1–37 (2023).

This review covers the dual protective and detrimental roles of microglia in various neurodegenerative diseases like Alzheimer's, Parkinson's, ALS, and others. It details how microglia interact with disease-specific proteins and contribute to pathology through impaired phagocytosis and neuroinflammation. The authors also discuss potential therapeutic strategies targeting microglia, such as enhancing phagocytosis, reducing inflammation, and promoting a neuroprotective phenotype. The review highlights how new single-cell technologies have improved understanding of microglial heterogeneity in these diseases. Overall, the authors conclude that modulating microglial functions is a promising avenue for developing new treatments for neurodegeneration.

- Microglia subtypes derived from scRNA seq

Kracht, L. et al. Human fetal microglia acquire homeostatic immune-sensing properties early in development. Science 369, 530–537 (2020).

This study investigated the development and maturation of microglia during human fetal development from 9 to 18 gestational weeks. Using single-cell RNA sequencing and chromatin accessibility analysis, the authors found that microglia undergo significant changes during this period, transitioning towards a more mature, immune-sensing state. The microglia started out heterogeneous but progressively acquired adult-like characteristics for immune surveillance in the CNS. This study suggests the developing fetal CNS may be vulnerable to environmental factors impacting microglia development. Additionally, disrupted microglia function could have potential implications for neurodevelopmental disorders.

- Hammond, T. R. *et al.* Single cell RNA sequencing of microglia throughout the mouse lifespan and in the injured brain reveals complex cell-state changes. *Immunity* **50**, 253-271.e6 (2019).

Hammond et al. used single-cell RNA sequencing to analyze microglia across the mouse lifespan and in brain injury. A key finding includes the identification of at least 9 distinct microglial states with unique gene expressions that change with age and brain injury. Microglial diversity peaks in young mice, and decreases with age until perturbed by injury, resulting in reactive subtypes. One reactive subtype matched human multiple sclerosis lesions, suggesting conserved microglial injury responses across species. The study mapped distinct microglial populations in the brain and linked findings in mice to human neurodegenerative diseases like multiple sclerosis, providing insights into potential microglial therapeutic targeting.

Limitations of previous work

Our work is inspired by Buckley et al. 2023[6]. One limitation of their work was that while their clocks showed high performance, the accuracy varied a lot across different cell types, possibly because some cell types might be better at predicting age than others. Microglia is one such cell type. We aim to hone in on only one cell type, microglia. Microglia (and possibly other cell types) tend to display a high degree of variability in terms of gene expression, morphology, and function throughout different ages starting from neurodevelopment until old age; this is something that this paper does not fully incorporate since they only use mice aged 3 to 29 months old.

Statement of Contributions

In this paper, we strived to build predictive models that can estimate the biological age of cells, specifically microglia, based on molecular markers, in this case gene expression.

Our contributions include:

- Identifying distinct microglia subtypes displaying unique gene expression patterns.
- Using machine learning algorithms (such as Random Forest and Elastic Net Regression) to probe into features and gene expression patterns of microglia subtypes that can accurately predict chronological age in both mice and humans.
- Establishing correlations between single-cell and bulk RNA seq data via pseudo bulk models to expand the scope of our study by using more prevalent bulk RNA seq datasets.
- Observing discrepancies between predicted biological age obtained through our model and chronological age, a phenomenon that could have implications for assessing individual differences in aging trajectories as well as differences between healthy individuals and individuals suffering from neurological conditions.

Methods

Datasets

All of our datasets are gene expression data from microglia.

- Single-cell RNA seq data from human fetuses aged 9 to 18 gestational weeks[7].
  - This single-cell dataset consisted of 15,782 cells with 977 measured genes.
- Single-cell RNA seq data from mice aged embryonic day 14.5 to postnatal day 540[8].
  - This single-cell dataset consisted of 35,450 cells with 2000 measured genes.
- Bulk RNA seq data from human fetuses aged 14 to 23 gestational weeks[9].
- Bulk RNA Seq data from male mice aged postnatal days 9 and 200[10]. Some mice were exposed to early life stress (ELS) while some were part of the control group (CTR). The mice were also either stimulated with lipopolysaccharide (LPS), with phosphate-buffered saline (PBS), or received no stimulation (Basal).

Description of methods

- Data pre-processing:
  After converting our scRNA seq data into an AnnData object, we did normalization (standard scaler) and log-transformation and retained 2000 highly variable genes.
- Unsupervised clustering on scRNA seq datasets:
  We applied the Leiden algorithm to identify distinct cell clusters. We determined the optimum number of nearest neighbors and principal components through scree plots and modularity score computation.
- Differential gene expression analysis:

Using Scanpy, we performed differential gene expression analysis with the Wilcoxon rank sum test to identify the top highly expressed genes in each cluster.

- Random forest to predict age
  We engineered frequency and functional features and trained a random forest model on 30 % of the samples and tested on the remaining 70%. For the human fetal dataset, we had replicated samples per fetus, so we trained on 30% and tested on 70% of the fetuses instead. We generated accuracy scores by comparing the predicted age values to the true values.
- Calculating Pseudobulk from scRNA seq
  We first created a pseudobulk representation of the single cell data to convert the data into a sample by gene format. To do this, the mean vector of gene expressions was computed per cluster. As an example, when this was performed on the mice single-cell dataset, the resulting output matrix had 48 rows each representing a mice sample and 42,000 columns, as gene expressions for each feature in the single-cell dataset across the 21 clusters were computed (21 * 2000). A pseudobulk data set was generated from both the mice and human fetus single-cell datasets. These datasets were then compared with the respective mice and human fetus bulk RNA seq data to verify whether one could extrapolate between both datasets and draw conclusive results about predicting age.
- PCA to assess any potential trends
  We first performed 2D PCA on our mice and human fetus data and color-coded by age to uncover any initial trends in the data.
- Linear regression analysis
  We trained a linear regression model with lasso on 30% of the pseudobulk data and tested on 70% and evaluated performance on the testing data using mean squared error.
- Elastic Net
  Because linear regression with the mice pseudobulk model didn't do too well at predicting age, we wanted to investigate more ways to test whether we can use pseudobulk, this time with the bulk datasets, to predict age.

  We trained an elastic net model on the mice and human fetus pseudobulk data and tested on the mice and human fetus bulk data. This was done in two approaches. For the first approach, we trained a model on a version of the pseudobulk dataset that pooled genes across all clusters by taking the average. For the second approach, we trained a model on each cluster rather than averaging gene expressions across all clusters. Regardless of the approach taken, the bulk dataset remained the same.

  For the mice dataset, box plots were generated for the pooled across clusters and cluster by cluster approaches to visualize how accurately the model could classify a bulk data sample as postnatal day 9 v.s. postnatal day 200.

  For the human fetus dataset, a multi-class regression approach was taken, so scatter plots were generated for the pooled across clusters and cluster by cluster approaches to visualize if there was any correlation between the predicted gestational age and true gestational age in the bulk data.

Schematic illustration of our methods

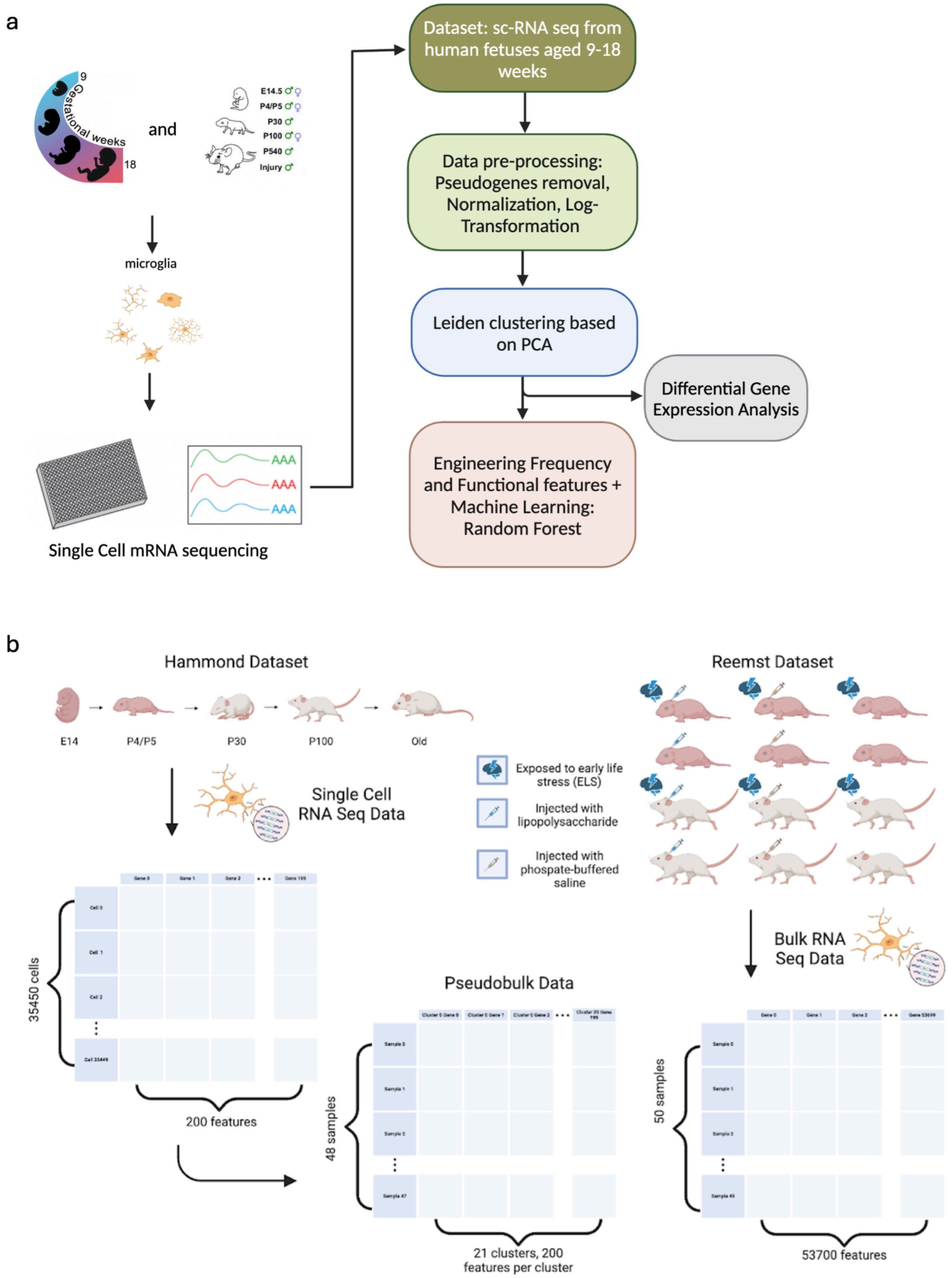

Figure 1: Schematic illustration of our methods. (a) Outline of our method for processing the single-cell RNA sequencing datasets from human fetuses and mice for unsupervised clustering, differential expression analysis, and age prediction with Random Forest. (b) Computational schema of how we extrapolated from two modalities, scRNA seq, and bulk RNA seq, shown here for the mice dataset.

Results

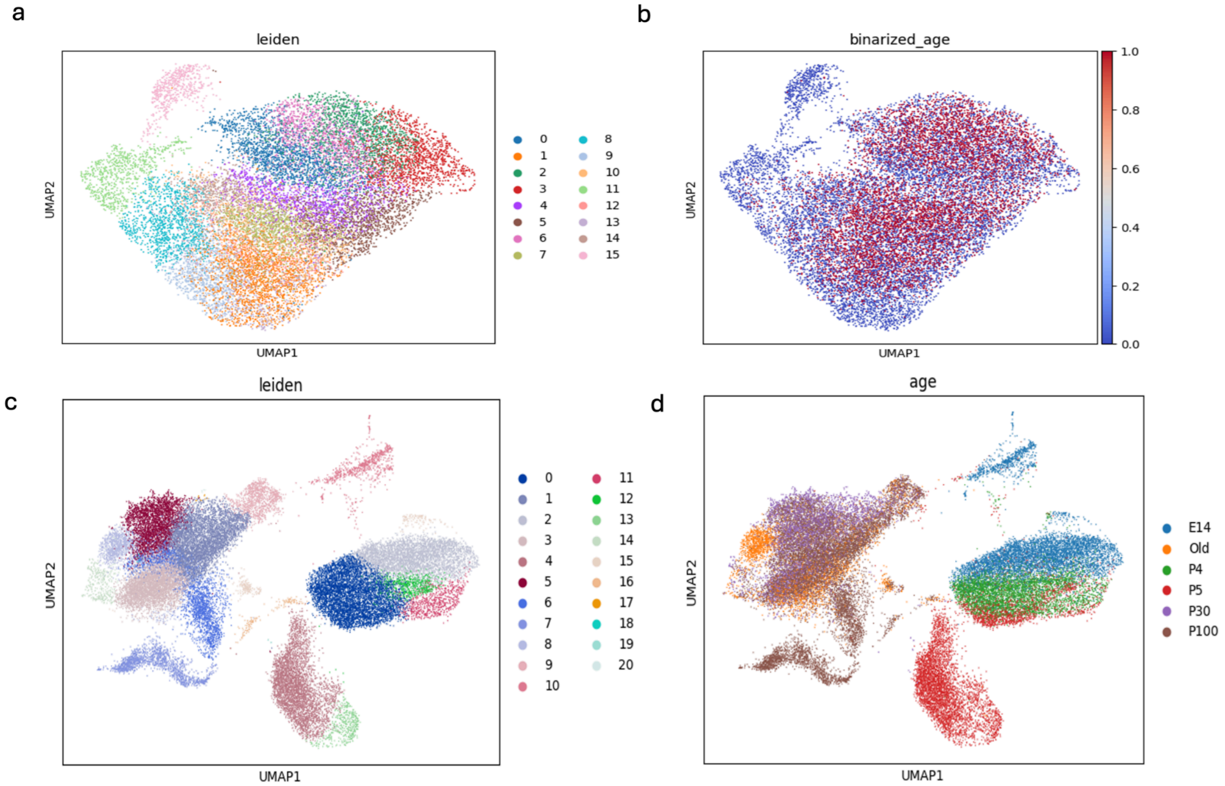

Figure 2. UMAPs of unsupervised Leiden clustering on two single-cell RNA seq datasets. (a) UMAP of human fetal dataset colored by all 16 Leiden clusters. (b) UMAP of human fetal dataset colored by age in trimester of pregnancy. '0' in blue indicates the first trimester of pregnancy, in this case, 9 to 12 weeks old, and '1' in red indicates the second trimester of pregnancy, here 13 to 18 weeks old. (c) UMAP of mouse dataset colored by all 21 Leiden clusters. (d) UMAP of mouse dataset colored by age. E14 is for embryonic day 14; P4, P5, P30, and P100 are for postnatal days 4, 5, 30, and 100 respectively and old is for postnatal day 540.

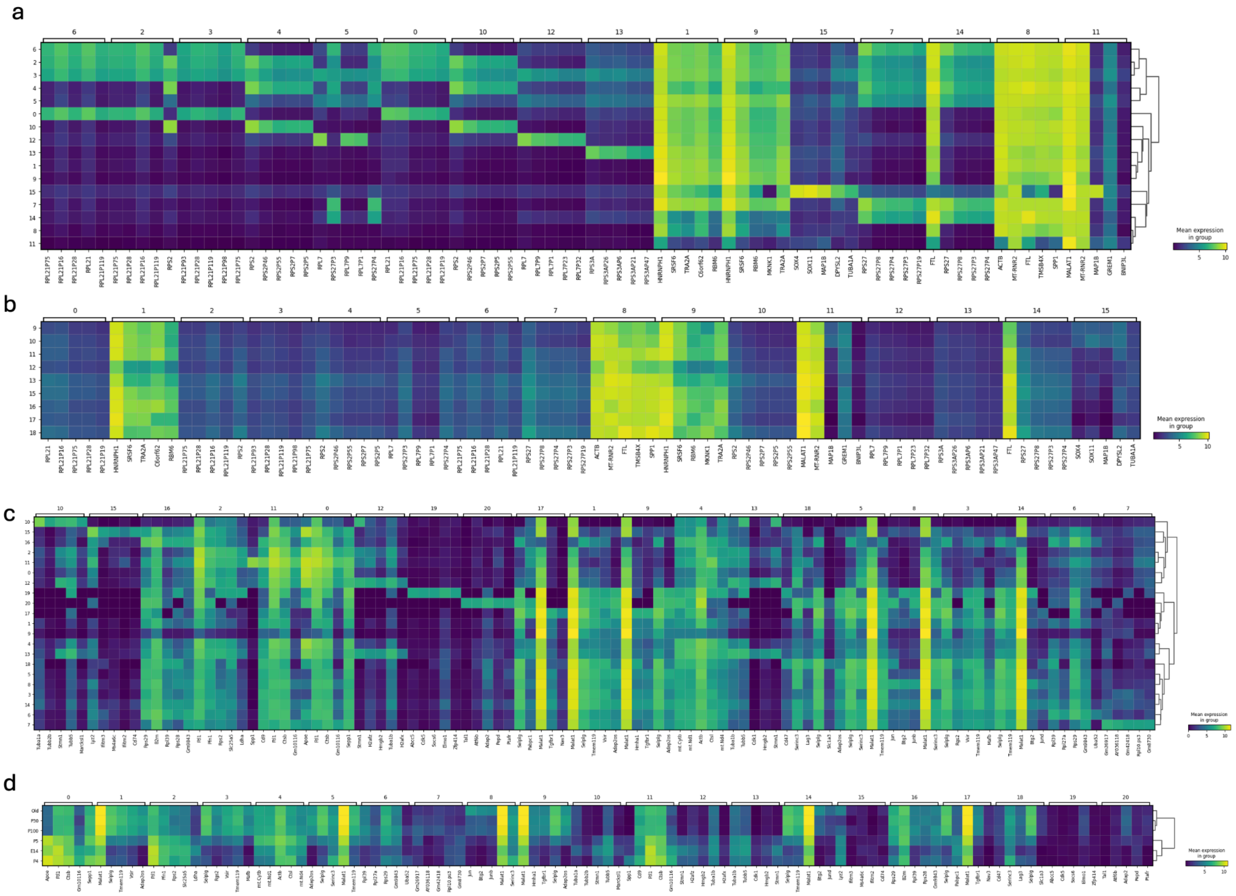

Figure 3. Matrix plots showing the top five genes per Leiden cluster determined from differential gene expression analysis on both datasets. (a) and (b): Human fetal data; gene expression of top five genes (a) compared to all clusters and (b) across gestational age. (c) and (d) Mouse data; gene expression of top five genes (c) compared to all clusters and (d) across mouse age.

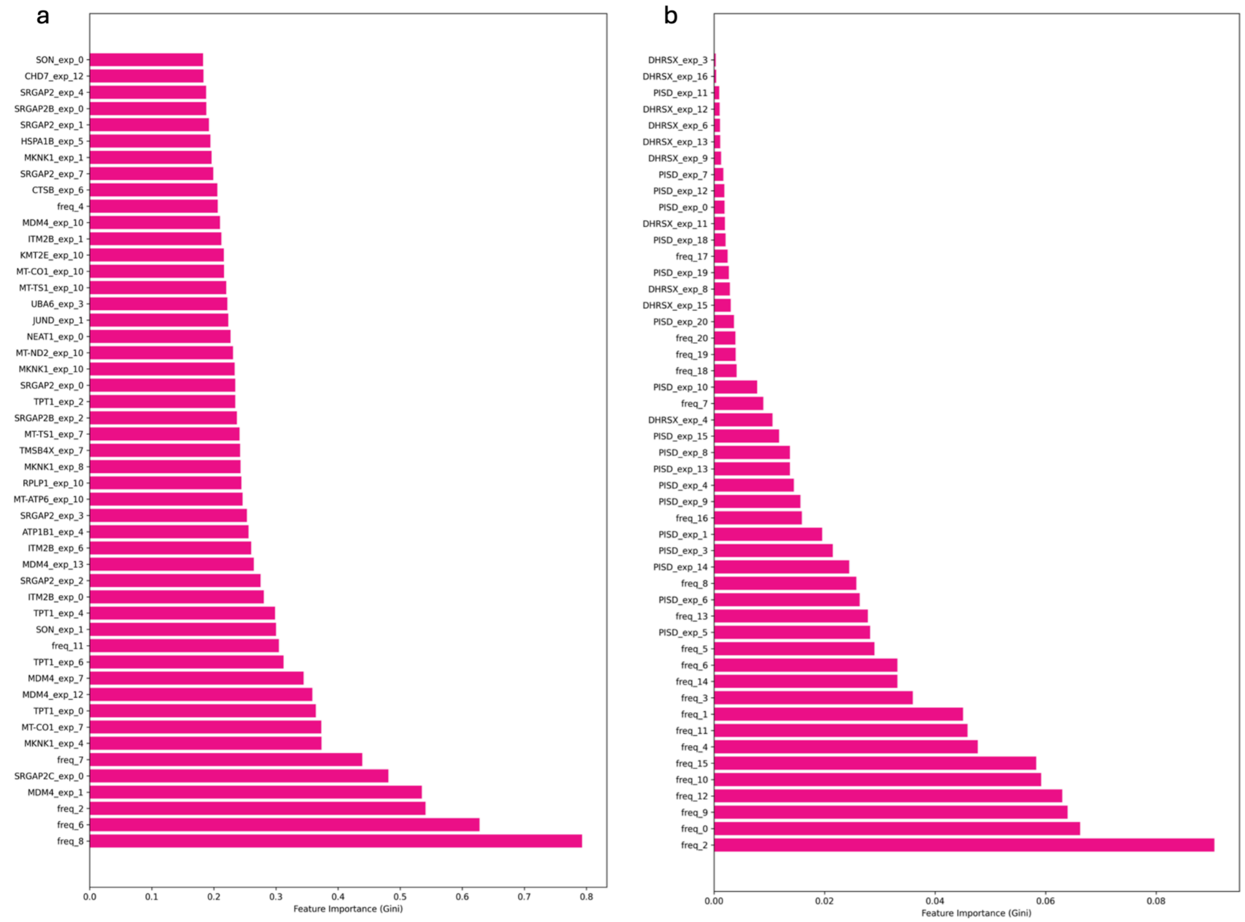

Figure 4. Gini importance plots illustrating the significant features in predicting age in both datasets using Random Forest classifier. (a) Top features in human fetal dataset. (b) Top features in mouse dataset.

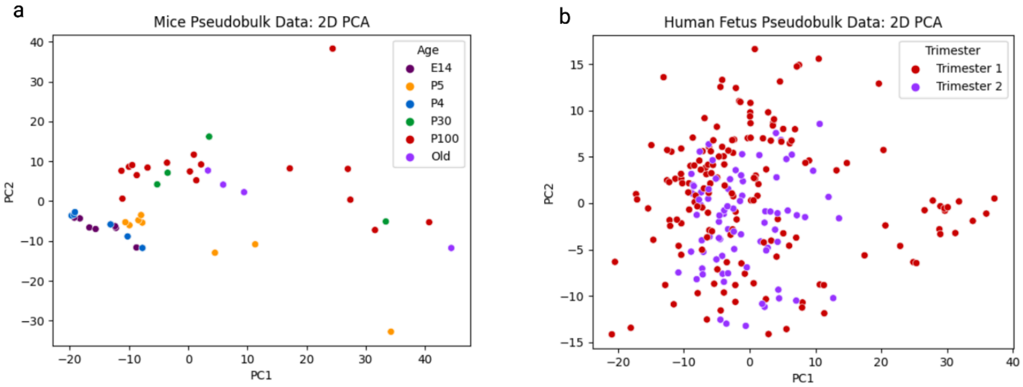

Figure 5: 2D PCA done on mice pseudobulk data, color-coded by age group (a) and human fetus pseudobulk data, color-coded by trimester (b).

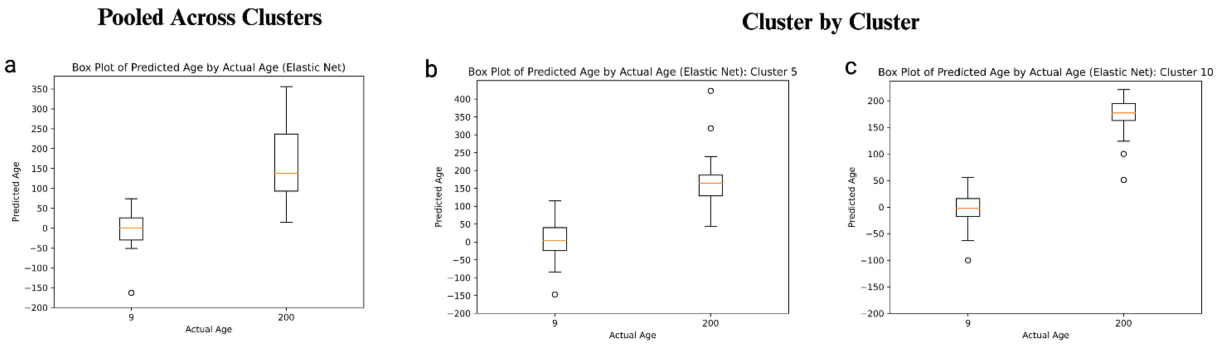

Figure 6: Predicted v.s. actual age across P9 and P200 mice in the bulk dataset. This was based on an Elastic Net model trained on gene expressions and mapped age values in pseudobulk data, where gene expressions were pooled across all 21 clusters (a), and on an Elastic Net model trained on gene expressions and mapped age values in the pseudobulk data for each cluster and tested on gene expressions in the bulk dataset (b and c)

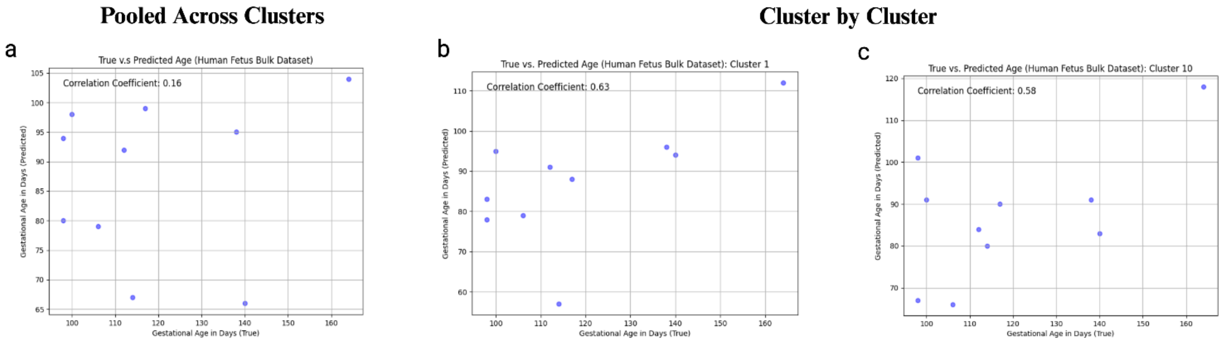

Figure 7: Predicted v.s. True gestational age across human fetuses in the bulk dataset. This was based on an Elastic Net model trained on gene expressions and mapped age values in pseudobulk data, where gene expressions were pooled across all 21 clusters (a), and on an Elastic Net model trained on gene expressions and mapped age values in the pseudobulk data for each cluster and tested on gene expressions in the bulk dataset (b and c)

Identifying microglia subtypes

Using unsupervised Leiden clustering, we were able to identify distinct groups of microglia in both the human fetal and mouse datasets. We found 16 clusters in the human fetal dataset (Fig. 2a) and 21 clusters in the mouse one (Fig. 2c). Upon coloring these clusters by age of sample, in the human fetal dataset, we were able to observe some clusters that belonged to the first trimester and second trimester of pregnancy, and others that showed mixed trimester membership (Fig. 2b). The cluster-age membership was more apparent in the mouse dataset where P30, P100, and Old clusters were grouped on the left of the UMAP, and E14, P4, and P5 were grouped on the right (Fig 2d).

A differential gene expression analysis using the Wilcoxon rank sum test allowed us to identify the top genes expressed in each microglia cluster. Each cluster had a different gene expression program (Fig. 3a and c) for both datasets. For instance, cluster 15 from the human fetal dataset corresponded to a cluster of cells belonging to the first trimester of pregnancy (Fig. 2a) and showed the highest expression of SOX4, SOX11, MAP1B, DPYSL2, and TUBA1A (Fig. 3a). Upon plotting gene expression by age, we found that three of these genes, SOX4, SOX11, and MAP1B, were indeed expressed highly at gestational weeks 9 and 10 (Fig. 3b).

Similarly, in the mouse dataset, cluster 0 coincided with cells aged E14, P4 and P5 (very young mice) (Fig. 2c.). The five highly expressed genes for this cluster were APOE, FTL1, CTSB, GM101106, and SEPP1, though interestingly some of these were also highly expressed in other clusters (e.g. both FTL1 and CTSB are highly expressed in cluster 11 which corresponds to cells from Old mice). When assessing the expression of these genes by age, we noted that APOE and GM101106 were indeed highly expressed in young mice (E14, P4, and P5); FTL1, CTSB, and SEPP1 did have high expression in young mice, though the change in expression in older mice was not drastic (Fig. 3d).

Predicting age using frequency and functional features with microglia scRNA seq data

We used a random forest classifier to predict age based on frequency and functional functional features. Frequency features were the different Leiden clusters that we obtained from doing unsupervised clustering while functional features were a gene list involved in cell differentiation, dementia, microglia development and function, inflammation, metabolic processes, immune function and circadian processes.

The random forest classifier was able to predict age of the human fetal dataset, when age was binarized by trimester of pregnancy, to a mean accuracy of 86% for frequency features, 76% for functional features and 77% when frequency and functional features were combined. Figure 4a summarizes the top features driving the model in age prediction, and as expected different microglia subtypes or frequencies are among the most important features.

When we ran the same model on the mouse dataset, we obtained a mean accuracy of 81% for predicting age with frequency features. However, with functional features, the prediction accuracy was only about 54%. Taken together, the mean prediction accuracy of both sets of features was about 80%. Figure 4b summarizes the top features driving the age prediction model and it is even more clear that frequency features are responsible for driving accurate age predictions.

Testing Pseudobulk and Bulk RNA seq data to predict age

We computed the pseudobulk from the single-cell RNA sequencing data from both mice and human fetuses as described in figure 1b and observed the distribution of the two first principal components for both datasets (Fig. 5a and b). No obvious separation of formation of clusters by age/trimester was observed in either PCA plots, which prompted us to continue exploring to see if further analysis could uncover any patterns.

From the Elastic Net model on the mice dataset, we see that the median predicted age when pooled across all clusters (Fig. 6a) is roughly around the true age for the P9 age group, but the model underpredicts the median age for the P200 age group. However, on a cluster level, we see some clusters, such as cluster 5 (Fig. 6b) and cluster 10 (Fig. 6c) tend to perform better, as the median predicted age for the P9 group is roughly the same as in Fig. 6a, but the model does not underpredict as severely for the P200 age group. This indicates that the gene expressions in clusters 5 and 10 had a more distinctive difference between the two age groups, and these differences were also reflected among gene expressions in the bulk dataset.

The results of the Elastic Net model on the human fetus dataset show that there is almost no linear correlation between the true and predicted gestational age when gene expressions are pooled across clusters, as the correlation coefficient is close to 0 (Fig. 7a). However, similar to what we saw in the mice data, certain clusters tend to do better at predicting age. In this case, clusters 1 (Fig. 7b) and 10 (Fig. 7c) had much higher correlation coefficients than Fig. 7a, signifying a possible linear relationship between predicted and true gestational age. This indicates that the gene expression profiles in these two clusters best reflect what is seen in the bulk dataset.

Discussion

The primary goal of this study was to enhance our understanding of the role microglia play in the aging process by leveraging single-cell and bulk RNA sequencing data. Specifically, we aimed to construct predictive models capable of estimating the biological age of microglia based on gene expression profiles. By analyzing transcriptomic data from microglia isolated from both mice and humans across a wide range of ages, we identified distinct microglia subtypes exhibiting unique molecular signatures. We then employed machine learning algorithms, including Random Forest (directly on scRNA seq data) and Elastic Net Regression (on pseudobulk and bulk RNA seq data), to identify the gene expression patterns and features most predictive of chronological age.

Our findings reveal that microglia display an interesting heterogeneity (as reported previously in Kratch et al 2020 and Hammond et al 2019)[7,8] in their gene expression profiles, reflecting diverse functional states that change dynamically throughout one's lifespan. Our approach, combining single-cell and bulk RNA sequencing data, allowed us to achieve reasonably accurate age predictions. Unlike the work of Buckley et al. (2023), which constructed aging clocks using single-cell transcriptomics alone, our method leverages the complementary strengths of both single-cell and bulk data modalities, hopefully enabling a more comprehensive and robust understanding of microglial aging.

Despite our promising results, our study has certain limitations. The accuracy of age predictions varied across different microglia subtypes when we used different machine learning models and datasets, suggesting that some subtypes may require alternative modeling approaches or additional features to improve their predictive power. Furthermore, our current models are based on a limited age range and specific experimental conditions. For instance, our scRNA seq dataset for human fetuses included first and second-trimester cells while the bulk version included only second-trimester cells. Similarly, the mice bulk dataset only included two ages, P9 and P200. This could have potentially affected the prediction accuracy of our models and generalizability to other age groups or environmental contexts.

Future research should focus on expanding the datasets to encompass a broader range of ages and conditions, thereby enhancing the robustness and applicability of our models. We could also test different clustering methods and further refine our pseudobulk algorithm. Additionally, the integration of multi-omics data, such as proteomics and metabolomics, could provide valuable complementary information, further refining the predictive power of the aging clocks. Once these improvements are accomplished, we should investigate all the genes that our models predict as being essential for our microglia clock of aging.

Conclusion

The development of microglia aging clocks could represent a significant milestone in our understanding of aging at the cellular level, especially in the context of the brain's immune system. Our work sheds light on the complex mechanisms involved in aging from a microglia viewpoint and could eventually pave the way for early diagnosis and personalized treatment strategies for age-related neurological diseases. By accurately predicting the biological age of microglia, we can gain deeper insights into individual variations in aging trajectories, enabling earlier interventions and personalized therapeutic approaches.

Ultimately, we hope our work is a crucial step towards deciphering the complexities of aging and developing effective strategies to promote healthy brain aging.

References

1. Prinz, M., Jung, S. & Priller, J. Microglia Biology: One Century of Evolving Concepts. *Cell* **179**, 292–311 (2019).
2. Harry, G. J. Microglia During Development and Aging. *Pharmacol. Ther.* **139**, 313–326 (2013).
3. Gao, C., Jiang, J., Tan, Y. & Chen, S. Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. *Signal Transduct. Target. Ther.* **8**, 1–37 (2023).
4. Pascoal, T. A. *et al.* Microglial activation and tau propagate jointly across Braak stages. *Nat. Med.* **27**, 1592–1599 (2021).
5. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276-1290.e17 (2017).
6. Buckley, M. T. *et al.* Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. *Nat. Aging* **3**, 121–137 (2023).
7. Kracht, L. *et al.* Human fetal microglia acquire homeostatic immune-sensing properties early in development. *Science* **369**, 530–537 (2020).
8. Hammond, T. R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* **50**, 253-271.e6 (2019).
9. Thion, M. S. *et al.* Microbiome Influences Prenatal and Adult Microglia in a Sex-Specific Manner. *Cell* **172**, 500-516.e16 (2018).
10. Reemst, K. *et al.* Early-life stress lastingly impacts microglial transcriptome and function under basal and immune-challenged conditions. *Transl. Psychiatry* **12**, 1–13 (2022).