Pakistan Journal of Bioinformatics (PJB)
Volume 01, Issue 02, Year 2026 — ISSN: 2222-7628

Identification and Validation of Blood Based Differentially Expressed Genes in Alzheimer Disease Using Integrated Transcriptomic Analysis

Yashfeen Alamgir*, Hassan Tariq
Department of Computer Science, University of Agriculture Faisalabad, Pakistan
* yashfeenalamgir@gmail.com

Abstract

Alzheimer disease is a progressive neurodegenerative disorder marked by memory impairment and cognitive decline. Early diagnosis always remain challenging because many established biomarker approaches rely on specialized imaging or invasive cerebrospinal fluid sampling, while robust blood based molecular biomarkers remain under validated Henriksen et al. (2014). To address the reproducibility gap in blood transcriptomic biomarker research, we applied a discovery and validation strategy using two independent whole blood gene expression datasets from the Gene Expression Omnibus: GSE63060 as the discovery cohort and GSE63061 as an external validation cohort Yu et al. (2021). Differential expression analysis was conducted using the limma framework with multiple testing correction Ritchie et al. (2015). In the discovery cohort, differentially expressed 100 genes were identified, and 15 genes were validated consistently in the independent cohort. Visualization with a heatmap of top 10 DEGs, volcano plot, boxplot of a representative probe, and overlap analysis supported clear transcriptional differences between control and Alzheimer disease sample. Receiver operating characteristic analysis showed strong diagnostic performance for representative probes in the discovery cohort (ILMN_2097421 AUC 0.871 and ILMN_1784286 AUC 0.862), that is indicating good discrimination between controls and Alzheimer disease Robin et al. (2011). Overall, the overlap genes being validated provide a foundation for downstream functional studies and reproducible blood based molecular signals and clinical evaluation

Keywords: Alzheimer's disease, blood-based biomarkers, transcriptomics, differential gene expression, ROC analysis, GEO

1 Introduction

Alzheimer disease is the most common cause of dementia and is characterized by progressive cognitive decline and functional impairment. Early and accurate detection is critical for patient care and for enabling timely intervention in research and clinical trials Jack et al. (2018). While cerebrospinal fluid biomarkers and the neuroimaging markers have strong diagnostic utility, they face barriers such as cost, invasiveness, and limited accessibility, which restrict routine large scale screening Henriksen et al. (2014).

Blood based biomarkers provide a practical alternative due to minimally invasive sampling and feasibility for repeated measurements. Transcriptomic profiling of peripheral blood can reveal systemic molecular changes associated with neurodegeneration, immune activation, and disease related pathways. However, many transcriptomic biomarker studies report cohort specific signatures without external validation, which can reduce reproducibility and generalization Henriksen et al. (2014). To strengthen evidence, biomarker candidates should be validated and in one cohort and discovered in an independent cohort.

In this study, we used an integrated transcriptomic pipeline with differential expression analysis and independent validation using two GEO datasets, GSE63060 (discovery) and GSE63061 (validation) Barrett et al. (2013). We quantified differential expression using limma, visualized disease associated patterns using standard plots, evaluated gene overlap across datasets, and assessed the diagnostic performance of DEGs with receiver operating characteristic analysis for representative probes.

2 Materials and Methods

2.1 Data sources and study design

Gene expression datasets were obtained from the NCBI Gene Expression Omnibus (GEO) Barrett et al. (2013). GSE63060 was used as the discovery dataset and GSE63061 was used as an independent validation dataset. Both datasets contain whole blood gene expression measurements from Alzheimer disease cases and cognitively normal controls Yu et al. (2021).

2.2 Preprocessing and group assignment

Downstream analysis was done using normalized expression matrices provided with the GEO series. Samples were assigned to control groups or Alzheimer disease using the corresponding phenotype annotations. Only the samples with clear diagnostic labels were included.

2.3 Differential expression analysis

Differential expression analysis was performed using limma, which fits linear models to expression data and applies empirical Bayes moderation to improve variance estimates Smyth (2004). Multiple testing correction was performed using the Benjamini and Hochberg false discovery rate procedure Benjamini & Hochberg (1995) Thresholds were used to consider the genes significant reported in analysis: adjusted p value less than 0.05 and absolute log2 fold change at least 0.5.

2.4 Discovery and validation strategy

Significant genes identified in GSE63060 were tested for consistency in GSE63061 dataset. Genes significant in both datasets were considered validated biomarkers, supporting the reproducibility across cohorts.

2.5 Visualization and performance evaluation

Clustering of top genes was visualized using a heatmap, Differential expression patterns were summarized using a vlacano, To illustrate group differences a representative probe was visualized using a boxplot, Overlap between datasets was illustrated using a Venn diagram. Diagnostic performance of representative probes was evaluated using receiver operating characteristic curves (ROC) and area under the curve (AUC) values computed using pROC Robin et al. (2011).

3 Results

3.1 Differential expression in the discovery dataset

In the discovery dataset GSE63060, DEGs analysis identified 100 statistically significant differentially expressed genes between control samples and Alzheimer disease under the reported thresholds. The volcano plot illustrates the distribution of effects and significance, highlighting genes meeting fold change and adjusted p value criteria.

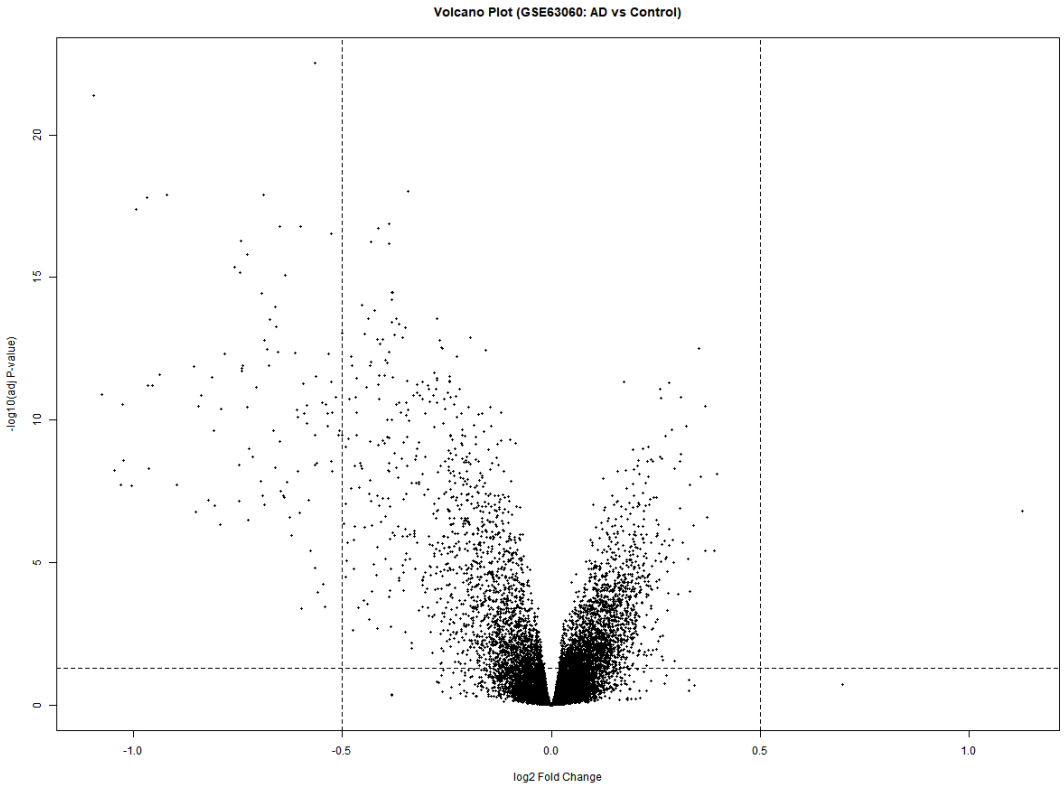

Figure 1: Volcano plot for GSE63060 (Alzheimer disease vs control).

3.2 Expression patterns of top differentially expressed genes

A heatmap of the top 10 differentially expressed probes revealed distinct expression patterns across samples, supporting disease associated transcriptional structure in whole blood. The top genes collectively showed separation trends between Alzheimer disease and control samples.

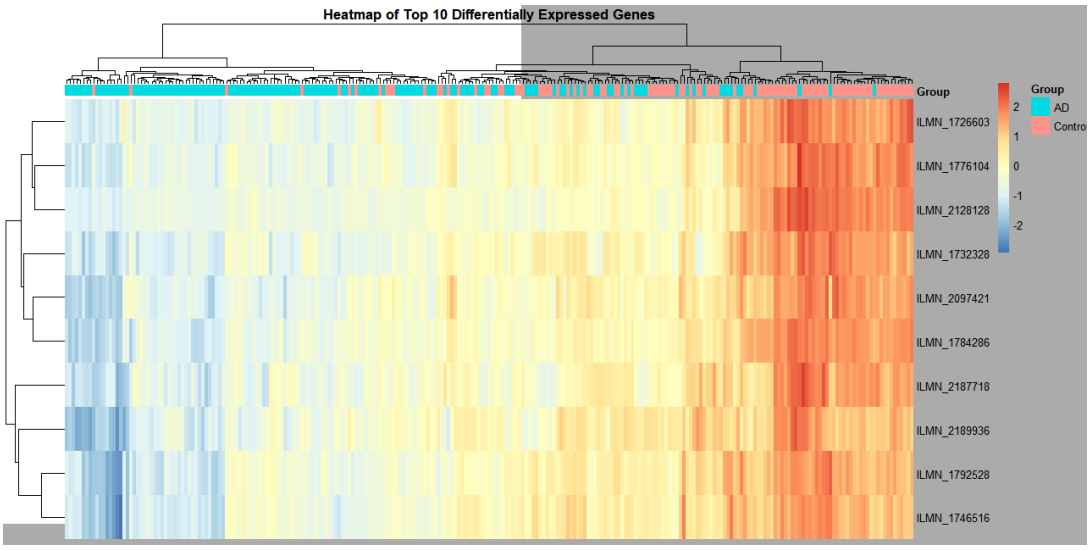

Figure 2: Heatmap of top 10 differentially expressed probes in GSE63060.

3.3 Representative probe level differences

A representative probe (ILMN_2097421) demonstrated clear expression differences between groups in GSE63060. The boxplot indicates lower expression in Alzheimer disease relative to controls, consistent with downregulation for this probe in the discovery cohort.

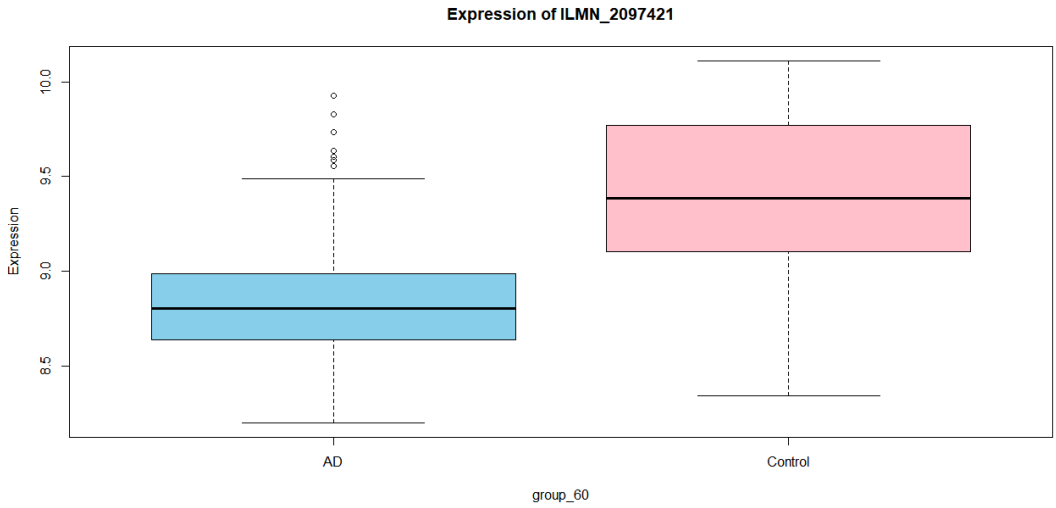

Figure 3: Boxplot of representative probe ILMN_2097421 in GSE63060 (Alzheimer disease vs control).

3.4 Independent validation and overlap between cohorts

Overlap analysis between GSE63060 and GSE63061 showed 15 genes significant in both datasets, indicating reproducible differential expression signals. The Venn diagram also shows genes unique to each dataset, which is expected due to cohort and technical variability, but the shared set provides stronger evidence for robust biomarker candidates.

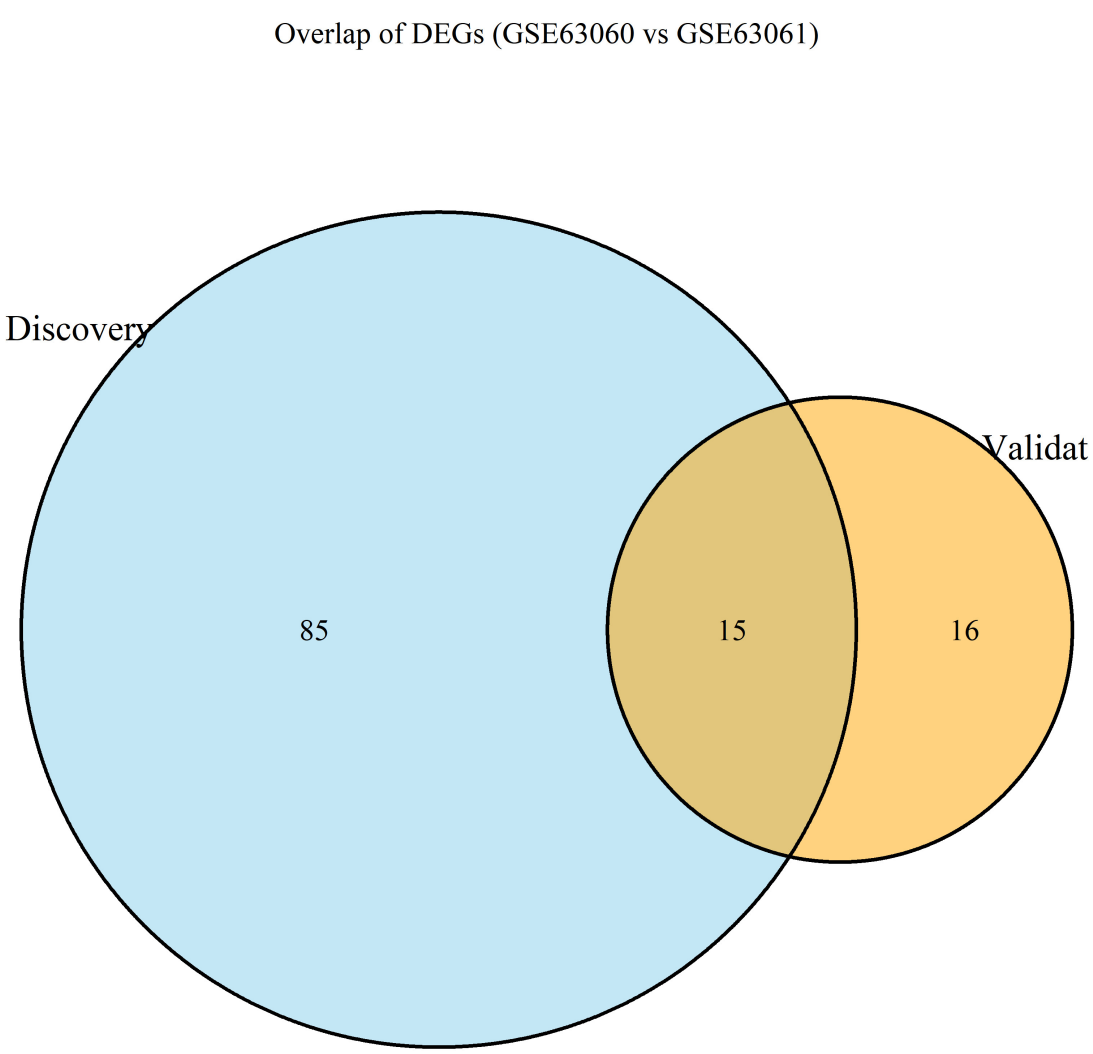

Figure 4: Overlap of significant differentially expressed genes between discovery (GSE63060) and validation (GSE63061).

3.5 ROC analysis in the discovery cohort

Receiver operating characteristic analysis in GSE63060 demonstrated good discriminatory performance for representative probes. ILMN_2097421 achieved an AUC of 0.871 and ILMN_1784286 achieved an AUC of 0.862, indicating strong separation between Alzheimer disease and control samples for these probes in the discovery dataset.

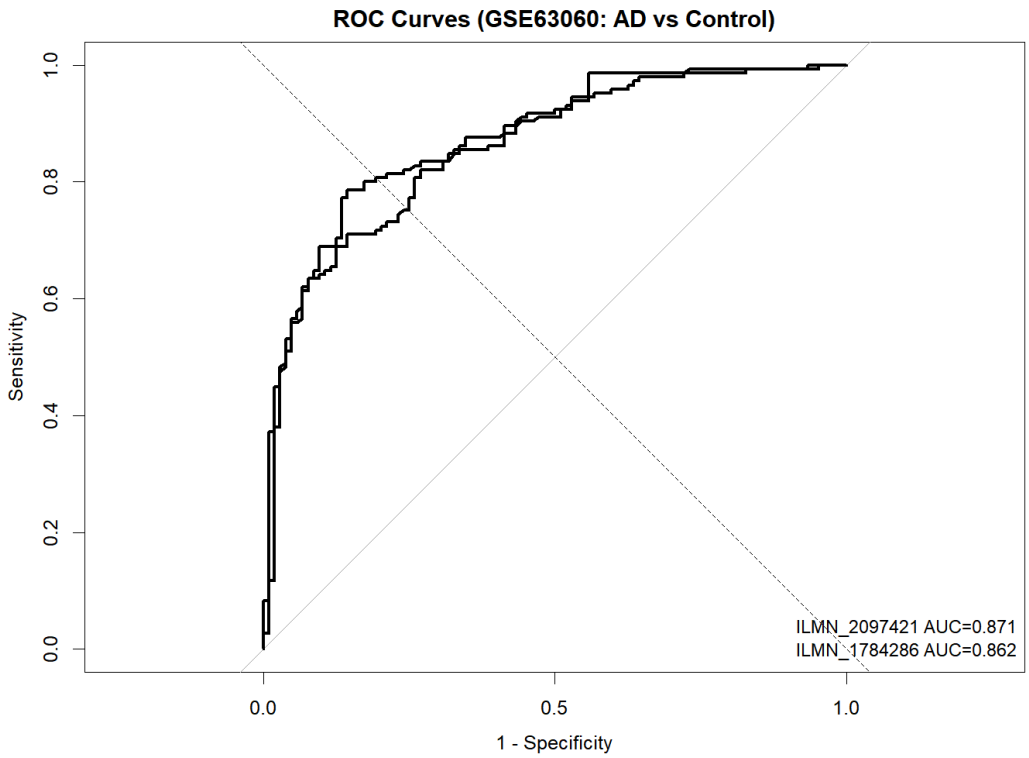

Figure 5: ROC curves in the discovery cohort (GSE63060) for representative probes (AUC values shown in the figure).

Table 1: Top 10 candidate blood based biomarkers identified from the GSE63060 discovery dataset.

Sr No	Rank	$\log_2$FC	AveExpr	t	P.Value	adj.P.Val	B
1	ILMN_2097421	-0.5662507	9.07901551	-12.404978	7.76E-28	2.97E-23	52.5066559
2	ILMN_1784286	-1.0964396	10.9548917	-11.969051	2.21E-26	4.24E-22	49.2138932
3	ILMN_2189936	-0.690057	12.3368815	-10.793945	1.56E-22	1.23E-18	40.5066062
4	ILMN_1776104	-0.920081	9.39844862	-10.78948	1.61E-22	1.23E-18	40.4740761
5	ILMN_1732328	-0.9689177	11.0244122	-10.72936	2.51E-22	1.61E-18	40.0366113
6	ILMN_2128128	-0.993525	9.07273246	-10.579441	7.61E-22	4.16E-18	38.9495873
7	ILMN_1792528	-0.601309	13.7324912	-10.361185	3.77E-21	1.61E-17	37.3773504
8	ILMN_1746516	-0.6505208	13.6976334	-10.343984	4.27E-21	1.64E-17	37.2539717
9	ILMN_2187718	-0.5273567	9.65359336	-10.238129	9.25E-21	2.95E-17	36.4965151
10	ILMN_1726603	-0.7437254	10.5736257	-10.146445	1.80E-20	5.30E-17	35.8429756

Table 2: ROC AUC values for representative probes in the GSE63060 discovery cohort (from ROC figure).

Probe ID	AUC
ILMN_2097421	0.871
ILMN_1784286	0.862

4 Discussion

This study applied a discovery–validation strategy to identify blood-based transcriptomic biomarkers for Alzheimer's disease. The key strength of this approach lies in the independent validation of candidate genes, as 15 biomarkers replicated across cohorts, directly addressing a common reproducibility limitation in transcriptomic biomarker research Henriksen et al. (2014). Visualization analyses demonstrated consistent disease-associated expression differences, while probe-level ROC analysis showed that individual probes achieved good classification performance in the discovery cohort.

Although the number of overlapping genes was smaller than the initial discovery set, this outcome commonly occurs in blood-based transcriptomic studies due to cohort heterogeneity, platform-specific effects, and biological variability. Importantly, replicated genes represent more reliable candidates for downstream biological interpretation and assay development. Future studies should map probes to gene symbols, investigate associated biological pathways, assess potential confounding factors, and evaluate classification performance in larger multi-cohort settings using multigene models and cross-validation strategies.

Study Limitations

This study is limited by the use of microarray-based transcriptomic data and lack of experimental validation. Future studies should include RNA-seq datasets, larger cohorts, and wet-lab validation to further confirm the diagnostic utility of the identified biomarkers.

Data Availability

The datasets analyzed in this study are publicly available from the NCBI Gene Expression Omnibus under accession numbers GSE63060 and GSE63061.

5 Conclusion

Using an integrated transcriptomic pipeline, we identified 100 differentially expressed genes in the discovery dataset and validated 15 genes in an independent cohort, supporting reproducible blood based molecular signatures of Alzheimer disease. Representative probes showed clear group differences and good ROC performance in the discovery cohort. These findings provide candidate biomarkers for future functional validation and clinical translation.

References

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L. & Edgar, R. (2013), 'Ncbi geo: Archive for functional genomics data sets—update', *Nucleic Acids Research* **41**(D1), D991–D995.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B (Methodological)* **57**(1), 289–300.

Henriksen, K., O'Bryant, S. E., Hampel, H., Trojanowski, J. Q., Montine, T. J., Jeromin, A., Blennow, K., Lönneborg, A., Wyss-Coray, T. & Soares, H. (2014), 'The future of blood-based biomarkers for alzheimer's disease', *Alzheimer's & Dementia* **10**(1), 115–131.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M. & Sperling, R. (2018), 'Nia-aa research framework: Toward a biological definition of alzheimer's disease', *Alzheimer's & Dementia* **14**(4), 535–562.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. (2015), 'Limma powers differential expression analyses for rna-sequencing and microarray studies', *Nucleic Acids Research* **43**(7), e47.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Muller, M. (2011), 'proc: An open-source package for r and s+ to analyze and compare roc curves', *BMC Bioinformatics* **12**, 77.

Smyth, G. K. (2004), 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical Applications in Genetics and Molecular Biology* **3**(1), Article 3.

Yu, W., Yu, W., Yang, Y. & Lü, Y. (2021), 'Exploring the key genes and identification of potential diagnosis biomarkers in alzheimer's disease using bioinformatics analysis', *Frontiers in Aging Neuroscience* **13**, 602781.