


Article

Computational Analysis of Marker Genes in Alzheimer's Disease across Multiple Brain Regions

Panagiotis Karanikolaos, Marios G. Krokidis *, Themis P. Exarchos and Panagiotis Vlamos 

Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, 49100 Corfu, Greece; pnskaranik@gmail.com (P.K.); exarchos@ionio.gr (T.P.E.); vlamos@ionio.gr (P.V.)

* Correspondence: mkrokidis@ionio.gr

Abstract: Alzheimer's disease (AD) is the most common cause of neurodegenerative dementia in the elderly, which is characterized by progressive cognitive impairment. Herein, we undertake a sophisticated computational analysis by integrating single-cell RNA sequencing (scRNA-seq) data from multiple brain regions significantly affected by the disease, including the entorhinal cortex, prefrontal cortex, superior frontal gyrus, and superior parietal lobe. Our pipeline combines datasets derived from the aforementioned tissues into a unified analysis framework, facilitating cross-regional comparisons to provide a holistic view of the impact of the disease on the cellular and molecular landscape of the brain. We employed advanced computational techniques such as batch effect correction, normalization, dimensionality reduction, clustering, and visualization to explore cellular heterogeneity and gene expression patterns across these regions. Our findings suggest that enabling the integration of data from multiple batches can significantly enhance our understanding of AD complexity, thereby identifying key molecular targets for potential therapeutic intervention. This study established a precedent for future research by demonstrating how existing data can be reanalysed in a coherent manner to elucidate the systemic nature of the disease and inform the development of more effective diagnostic tools and targeted therapies.

Keywords: big data; transcriptomics; dimensionality reduction; brain; Alzheimer's disease



Citation: Karanikolaos, P.; Krokidis, M.G.; Exarchos, T.P.; Vlamos, P.

Computational Analysis of Marker Genes in Alzheimer's Disease across Multiple Brain Regions. *Information* **2024**, *15*, 523. <https://doi.org/10.3390/info15090523>

Academic Editors: Haifeng Wang, Norma B. Ojeda and Lu He

Received: 25 June 2024

Revised: 8 August 2024

Accepted: 26 August 2024

Published: 27 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Alzheimer's disease (AD), the most common form of dementia among older adults, is characterized by a progressive decline in cognitive function and the pathological accumulation of amyloid-beta plaques and tau tangles [1]. The disease's complexity is further compounded by its multifactorial nature, involving genetic, environmental, and molecular factors. Understanding the pathogenesis of AD at a cellular and molecular level has been a significant challenge, largely due to the heterogeneous nature of the brain and the intricate interactions between various cell types [2]. The considerable interval between the onset of initial pathophysiological changes and the emergence of clinical symptoms suggests an Alzheimer's disease continuum, encompassing various transitional stages. At the earliest point in this continuum is the preclinical AD phase. Following this is the prodromal stage known as mild cognitive impairment (MCI), characterized by cognitive deficits that do not significantly interfere with daily activities. Beyond MCI lies the dementia phase [3].

The development of single-cell RNA sequencing (scRNA-seq) has provided an unprecedented opportunity to explore multiple complexities at the resolution of individual cells [4,5]. scRNA-seq technology has emerged as the leading method for deciphering the diversity and intricacies of RNA transcripts at the individual cell level. It enables the exploration of the composition, functions, and heterogeneity within various organized tissues, organs, or organisms. The procedures of scRNA-seq primarily involve single-cell isolation and capture, cell lysis, reverse transcription (conversion of RNA into cDNA), cDNA amplification, and library preparation [6]. This technology allows detailed examination

of gene expression patterns within single cells, offering insights into the cellular composition of tissues and the distinct role of different cell types between healthy and diseased conditions. In the context of AD, scRNA-seq has the potential to reveal how different brain regions are uniquely affected by the disease, highlighting variations in cellular responses and molecular pathways [7,8]. Increased amyloid-beta secretion in AD olfactory mucosal cells and detailed cell-type-specific gene expression patterns have been reported through scRNA-seq as well as 240 differentially expressed disease-associated genes compared to the cognitively healthy controls and five distinct cell populations [9]. Specific transcriptional changes in different cell types such as neurons, astrocytes, and microglia from post-mortem human brain tissue of AD patients and control subjects have been identified, revealing distinct transcriptional alterations in these glial cells and suggesting their pivotal roles in the disease's progression [10].

In a previous work by our group, gene–gene interaction networks integrated with scRNA-seq expression profiles were constructed, while the most active subnetworks were isolated from the entire network topology [11]. Moreover, combining both deep learning and machine learning processes examining scRNA-seq data obtained from the peripheral blood of both AD patients with an amyloid-positive status and healthy controls with an amyloid-negative status, differentially expressed genes have been observed which were mainly enriched in the regulation of the immune system, interferon-gamma-mediated signalling, and the cellular defence response [12]. Drawing upon data from a database called scREAD (single-cell RNA-seq Database for Alzheimer's Disease), another study centred on astrocytes isolated from the entorhinal cortex of both AD patients and healthy individuals. The study identified differentially expressed genes and extracted disease-specific pathways and gene ontologies, along with predicting drugs and natural products capable of regulating AD-specific signatures in astrocytes [13]. Furthermore, disruptions in synaptic signalling and cell-cycle regulation across different cell types in the prefrontal cortex of AD patients have been observed, offering insights into neuronal dysfunction and degeneration mechanisms in the disease while critical pathways involved in synaptic signalling and cell-cycle regulation have been significantly disturbed in the prefrontal cortex, highlighting potential therapeutic targets [14]. Variations in immune response genes and disruptions in the insulin/IGF1 signalling pathways have also been identified, crucial for understanding the disease's early stages, pinpointing potential biomarkers for early detection and intervention, which could be pivotal in monitoring the disease's progression [15]. An upregulation of the insulin/IGF1 signalling machinery seems contrary to the notion of central insulin resistance. Alternatively, it might represent a compensatory mechanism that enhances neuroprotection in areas of the Alzheimer's disease brain that have not yet experienced neuronal loss [15]. Pathways and neurotransmitters involved in AD are summarized in Figure 1.

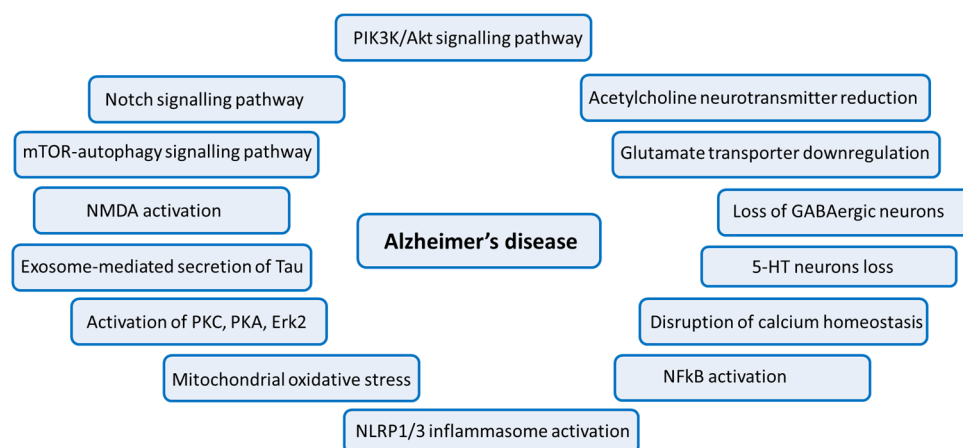


Figure 1. Pathways and neurotransmitters involved in AD.

The datasets utilized for the purposes of the present work were obtained from the scREAD database [16] which encompasses both scRNA-seq and snRNA-seq datasets derived from postmortem human brain tissue exhibiting AD and animal models with AD pathology. Control datasets sourced from healthy, non-AD samples were also included. By employing advanced computational methods, researchers can integrate scRNA-seq datasets from various brain regions to form a holistic view of the disease. In a recent work by our group, scRNA-seq data between the mice cortex and hippocampus from healthy and AD samples have been compared, and differentially expressed genes were observed, mainly enriched in muscarinic acetylcholine receptors, dopamine receptors, and perisynaptic extracellular matrix [17]. The present study leverages computational techniques to analyse scRNA-seq data from multiple brain regions impacted by AD and reinforce existing studies that AD manifests differently in different brain regions and cell types. RNA-seq can explore differential gene expression across multiple brain regions, providing new challenges to identify key biological processes from a molecular perspective. By synthesizing data across the entorhinal cortex, prefrontal cortex, superior frontal gyrus, and superior parietal lobe, we aimed to build a comprehensive model of AD's impact on the brain's cellular and molecular landscape. Through this approach, we emphasize the potential of computational analyses to deepen our understanding of neurodegenerative diseases such as AD. This allows for the comparison of cellular and molecular profiles across different areas, identifying common and distinct elements of the pathology.

2. Materials and Methods

2.1. scRNA-seq Data Collection

The present study used scRNA-seq datasets from the scREAD database [16], comprising cells from four distinct human brain regions to explore the cellular and molecular landscape of AD. More specifically, the analysis included selected datasets from the entorhinal cortex, prefrontal cortex, superior frontal gyrus, and superior parietal lobe, with individual datasets representing both healthy and AD-affected subjects as previously outlined. For the entorhinal cortex, we included data from healthy control subjects (datasets AD00201 and AD00202) and compared it to data from subjects with AD (datasets AD00203, AD00204, AD00205, and AD00206). In the case of the prefrontal cortex, the control group comprised datasets AD01101 and AD01102, while the disease group included AD01103 and AD01104. The superior frontal gyrus control group data were sourced from dataset AD00801, and the disease group data were represented by AD00802 and AD00803. Lastly, six datasets comprising cells from the superior parietal lobe were utilized; among these, two datasets corresponded to control (AD01201 and AD01202) while the remaining four represented cells with AD pathology (AD01203, AD01204, AD01205, and AD01206). Regarding the number of samples for each brain area: entorhinal cortex control: 27,892; entorhinal cortex disease: 42,635; prefrontal cortex control: 29,752; prefrontal cortex disease: 18,326; superior frontal gyrus control: 30,608; superior frontal gyrus disease: 50,256; superior parietal lobe control: 411; superior parietal lobe disease: 2486.

2.2. Determination of the Optimal Number of Principal Components for Data Analysis

The `determine_optimal_pcs` function helps to find the best number of principal components (PCs) for single-cell RNA sequencing data. This function uses the explained variance from Principal Component Analysis (PCA) to find the minimum number of components needed to explain a certain amount of total variance. This method avoids the manual and subjective elbow method, making the process more objective and reproducible. The function performs PCA on the dataset using the scanpy library's `sc.tl.pca` function. The number of components is set by `n_pcs`, and the variance explained by each component is saved in the `adata.uns['pca']['variance_ratio']` attribute. It then sums these variances cumulatively. Using `np.searchsorted`, the function finds the smallest number of PCs needed to exceed the `variance_threshold`. If the threshold is not reached within the computed

components, it uses all components. Otherwise, it returns the optimal number of PCs needed to surpass the threshold.

By automating the selection of the number of principal components, the function reduces the need for manual intervention and subjective judgment. Using a consistent and objective method for determining the optimal number of PCs ensures that results are reproducible across different datasets and analyses. Automating the process allows for quicker analysis, which is particularly beneficial when dealing with large datasets typical in single-cell RNA sequencing. Additionally, users can easily adjust the `n_pcs` and `variance_threshold` parameters to suit their specific needs and preferences.

2.3. Identification and Correction of Batch Effects in scRNA-seq Datasets

Batch effects were managed using AnnData objects to integrate smoothly into scRNA-seq analysis pipelines. Key steps included identifying, evaluating, and correcting batch effects to ensure reliable downstream analyses. Initially, checks were conducted to detect multiple batches, determining if batch correction was needed. If only a single batch was present, the function ceased to conserve resources. UMAP was used for initial visualization, followed by ANOVA to detect batch effects, with *p*-values adjusted via the Benjamini–Hochberg method. Correction methods, such as ‘combat’ and ‘harmony’, were offered for customized batch correction. Post-correction UMAP visualization confirmed the effectiveness by eliminating distinct batch clusters, ensuring data homogenization for subsequent analyses.

2.4. scRNA-seq Data Integration and Preprocessing Overview

To ensure the appropriate integration and traceability, each dataset was duplicated and labelled with its brain region of origin. These datasets were then merged into a unified AnnData object using the `sc.concat` function, preserving the variability and source of each data point. Gene filtering was performed to harmonize the gene set across all datasets, enabling reliable comparative analyses. Strict quality control (QC) metrics were applied to remove outliers, targeting cells with abnormal gene counts or high mitochondrial gene expression. Specifically, cells were filtered based on gene count (300 to 2500) and mitochondrial gene expression (<10%) to exclude non-viable cells and potential doublets, ensuring the dataset’s quality for accurate analysis.

2.5. Integration of Datasets from Various Brain Regions

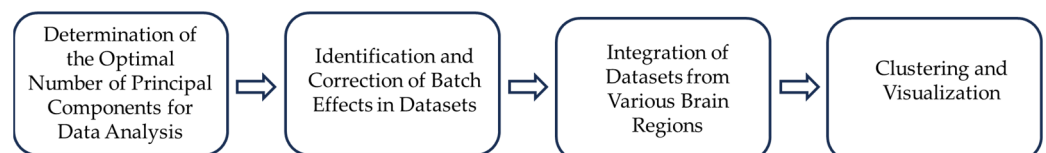
The process began with normalization and logarithmic transformation to standardize gene expression levels across cells, ensuring comparability. Normalization (using `sc.pp.normalize_total`) with a target sum of 1×10^4 and logarithmic transformation (via `sc.pp.log1p`) stabilized variance, making the data more homogeneous and ready for analysis by mitigating technical disparities. Next, highly variable genes were identified and selected to highlight biological signals amid technical noise, focusing on genes with significant variability. This step is crucial for understanding the underlying biology of the sample. Doublet detection and removal addressed artifacts from multiple cells being sequenced as one. The `remove_doublets` function, using the Scrublet tool, predicted and eliminated doublets from the dataset. Scrublet simulates doublet formation and assigns a doublet score to each cell. Parameters like `min_counts`, `min_cells`, `min_gene_variability_pctl`, and `n_prin_comps` were adjusted to fine-tune detection. Doublet scores and predictions were annotated within the AnnData object for further validation. Finally, cells predicted as doublets were filtered out, resulting in a purified dataset of putative singlets. This refined dataset, free of doublet artifacts, was ready for accurate and reliable downstream analyses.

2.6. Clustering and Visualization

The Leiden clustering algorithm (`sc.tl.leiden`) applied to the control scRNA-seq dataset, using a resolution of 0.5 and a random state of 5 for consistent clustering. This organized the cells into distinct clusters based on gene expression profiles, identifying unique cell

populations. Following clustering, Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction, projecting gene expression data into two dimensions. Clusters were colour-coded, and visualization enhancements included point outlines, an on-plot legend, and a custom colour palette, improving clarity. To understand the spatial distribution of clusters, a contingency table was constructed to examine their distribution across brain areas. Clusters with cells from three or more brain areas indicated broader representation and potential biological significance. Cells from these representative clusters were retained, and the Leiden algorithm was re-applied with the same parameters. This refined clustering accurately represented underlying biological diversity. The refined clusters were visualized again in UMAP space for comparison, confirming clustering robustness and providing deeper insights into cell population distribution. The Leiden algorithm was chosen for its effective handling of scRNA-seq data. This algorithm can explore clear and meaningful clusters, especially in large and complex datasets. It optimizes modularity better than other methods, which means the clusters it finds are strong and accurate. The Leiden algorithm's design helps it find smaller groups within the data, giving a detailed view of cellular differences. Additionally, it produces stable and repeatable results, which is important for consistent outcomes in different analyses, making our findings more reliable.

A differential gene expression analysis using the Wilcoxon rank-sum test (**sc.tl.rank_genes_groups**) identified key marker genes for each cluster. These markers were visualized using a dot plot, highlighting distinct gene expression patterns and facilitating specific cell type identification. Clusters were annotated with cell type identities based on marker gene profiles, including inhibitory neurons, microglia, astrocytes, and excitatory neurons. These annotations were visualized in a UMAP plot, colour-coded by cell type, depicting their distribution across the UMAP space. These comprehensive analyses and visualizations provided a deep understanding of the cellular heterogeneity and molecular characteristics within the control scRNA-seq dataset, offering valuable insights into the brain areas' cellular compositions and functional organizations, enhancing our understanding of the underlying biology. The followed pipeline is summarized in Scheme 1.



Scheme 1. scRNA-Seq data analysis flowchart used in this study.

3. Results

According to our analysis, findings are meticulously visualized through a series of tools designed to enhance our interpretative ability. This includes the ranking of genes associated with each cluster and the use of a heatmap, which illustrates the marker genes and their expression patterns across the clusters. Through these visualization techniques, we not only highlight the biological differences between distinct clusters but also underscore the potential discovery of unique cellular identities or states. This comprehensive approach significantly enriches our understanding of the dataset's underlying biology, paving the way for further explorations into the cellular intricacies of the brain. The series of graphs represent a comparative analysis of gene expression distributions within individual clusters against the backdrop of all other cells not included in those clusters (Figures 2 and S1). Each plot is labelled with a cell identifier signifying the specific cluster being analysed against the collective remainder of the dataset. In each subplot, the horizontal axis, marked as 'ranking', orders the genes from left to right based on their relative importance or impact within the cluster's gene expression profile. The 'score' on the vertical axis quantifies the level of differential expression, with higher scores potentially indicating a greater degree of upregulation within the cluster compared to the rest of the cells.

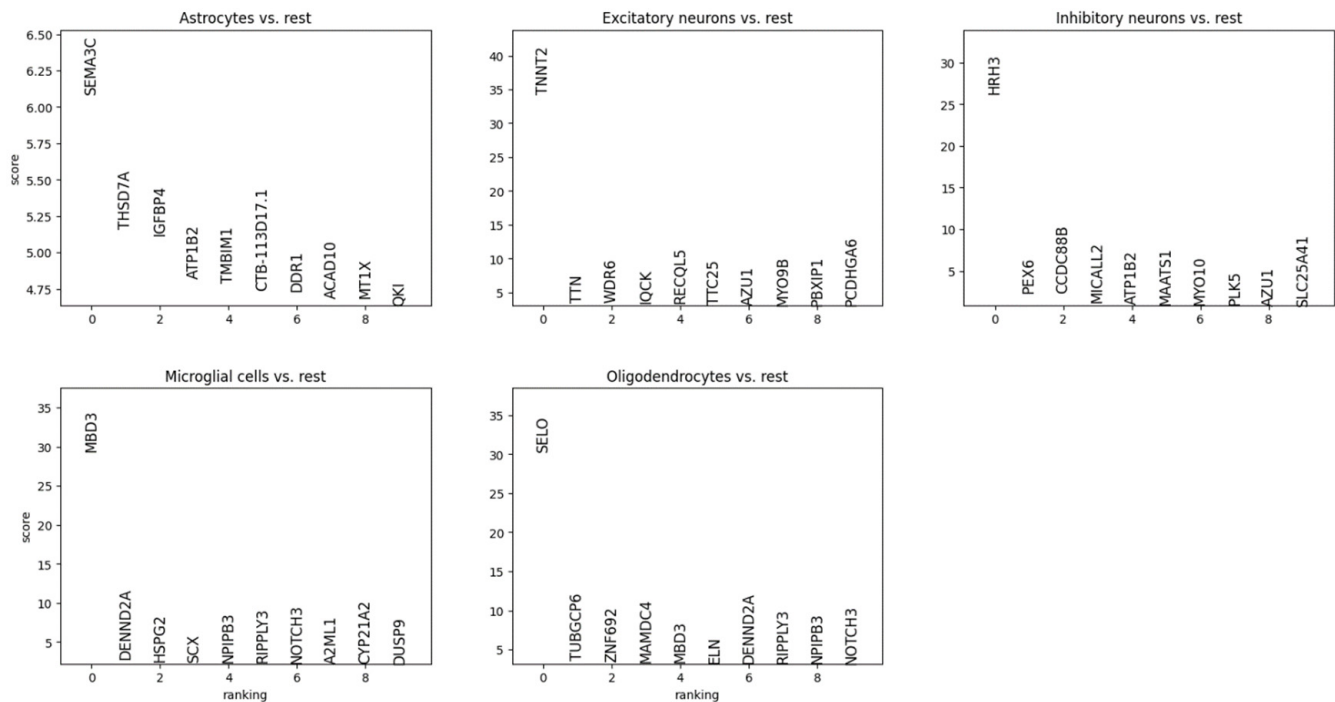


Figure 2. Differential expression of the top 10 ranked genes in each identified cell group compared to the rest, as analysed in the AD dataset.

Points plotted in each graph, as Figure 2 shows, represent individual genes, with their position reflecting their ranked relevance and expression score within the cluster. For instance, a gene that appears toward the left with a high score is of substantial significance within that cluster and shows a notably different expression level compared to cells outside of the cluster. The discrete distribution of points across the ranking spectrum allows us to discern which genes are most characteristic of each cluster. This comparative graphical approach is invaluable for highlighting the genes that distinguish each cluster from the rest, thereby providing a detailed view into the molecular identity of each cell population. By observing the patterns and positions of these genes across the series of plots, researchers can draw conclusions about the biological processes that may be predominant in each cluster and identify potential targets for further experimental investigation.

Furthermore, an intuitive graphical representation to visualize gene expression across distinct categories was performed using heatmaps, as Figure 3 illustrates. These values are systematically arranged and grouped by specific categories, providing a clear differentiation between different gene expression levels. Within this matrix plot, each column corresponds to a category or cluster, such as different cell types, tissues, or experimental conditions. For each cluster within the plot, the expression of genes is quantified in terms of fold change values as a measurement comparing the expression level of a gene in one condition to its level in a reference condition. A positive fold change value signifies that there is an upregulation or increase in the expression of the gene within that cluster, suggesting that the gene is more active compared to the reference condition. Conversely, a negative fold change indicates a downregulation or decrease in gene expression, implying that the gene is less active or potentially repressed in the compared condition. This differential expression analysis, highlighted through Figure 3, enables us to identify genes that show significant changes in expression across different categories or conditions, facilitating insights into biological processes and disease mechanisms.

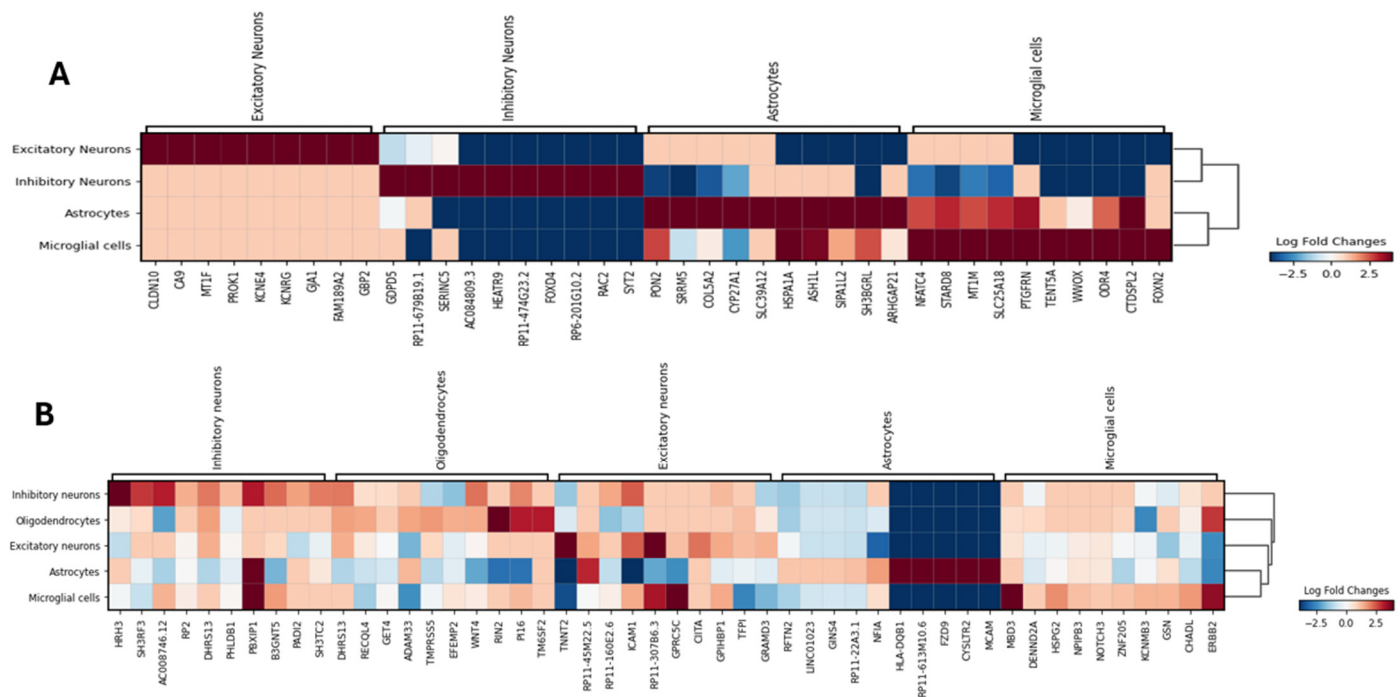


Figure 3. Heatmap of the log fold changes in gene expression (**A**) for the top 10 differentially expressed genes across four cell groups within the control dataset, (**B**) for the top 10 differentially expressed genes across five cell groups within the AD dataset.

Figure 4 shows dot blot for summarizing the expression of each gene across all cells within a group and visualizes scRNA-seq expression data across different clusters. Cell groups are shown along the horizontal axis, and genes are arranged along the vertical axis. The dendrogram at the top represents hierarchical clustering based on expression profiles, grouping similar expression patterns together. The size of each dot indicates the fraction of cells within a group expressing the gene marker (with the percentage scale shown in the top right corner), while the colour intensity represents the mean expression level of the gene in that specific group (as indicated by the colour scale at the bottom). This visualization facilitates the assessment of both the prevalence and intensity of gene expression across different groups, providing insights into the dynamics of gene expression and cellular diversity within the sample. In our analysis, the dot plot was standardized by variance to enable comparison across different genes. Additionally, dot plots, as illustrated in Figures S2 and S3, provide an insightful visualization of scRNA-seq expression data across various clusters, facilitating comparisons between control and disease groups. Figure S2 displays the expression patterns of key genes across different cell types within the control group, including inhibitory neurons, excitatory neurons, astrocytes, oligodendrocyte precursor cells, oligodendrocytes, and microglial cells. Figure S3 presents the dot plot for the disease group, highlighting alterations in gene expression patterns due to the pathological condition. Significant alterations in the expression levels and prevalence of specific genes across cell types are evident when compared to the control group. These dot plots are an invaluable tool for visualizing the complexity of gene expression across different cell populations, offering a thorough overview of cellular heterogeneity and the effects of disease on gene expression dynamics.

Furthermore, violin plots as Figure 5 shows, display the distribution of expression levels for the top five differentially expressed genes in each group compared to the rest of the groups in the control dataset. The width of each plot indicates the density of cells at different expression levels, while the split view highlights differences between the two conditions being compared.

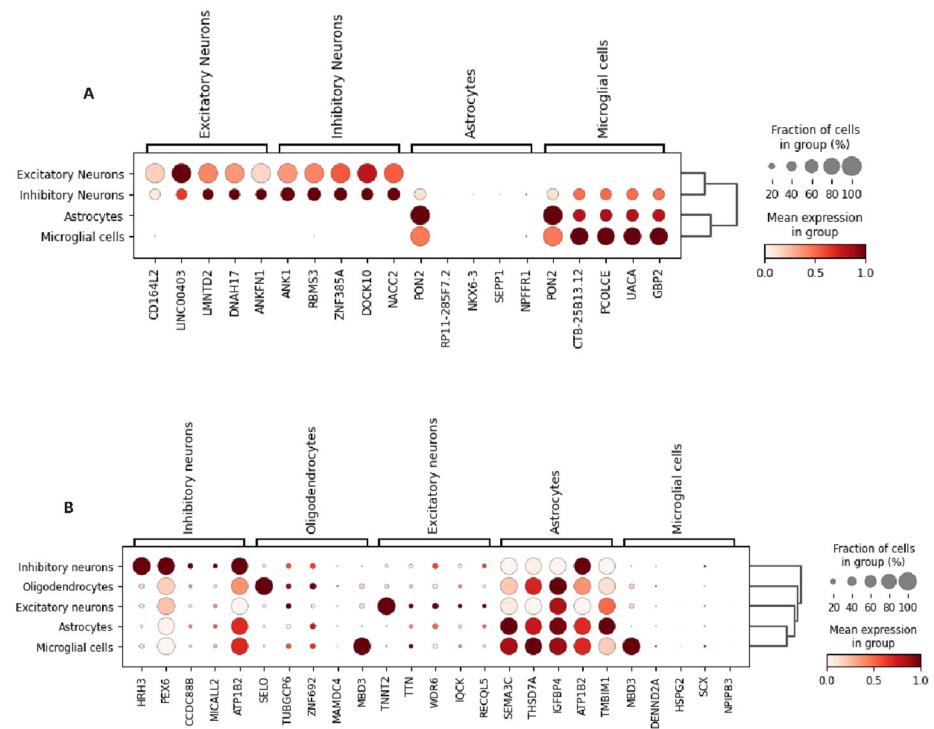


Figure 4. A dot plot visualization of scRNA-seq data. (A) Dot plot presents the differential expression of the top five genes within each of the four cell groups in the control dataset. (B) Dot plot depicts the differential expression of the top five genes within each of the cell groups in the AD dataset.

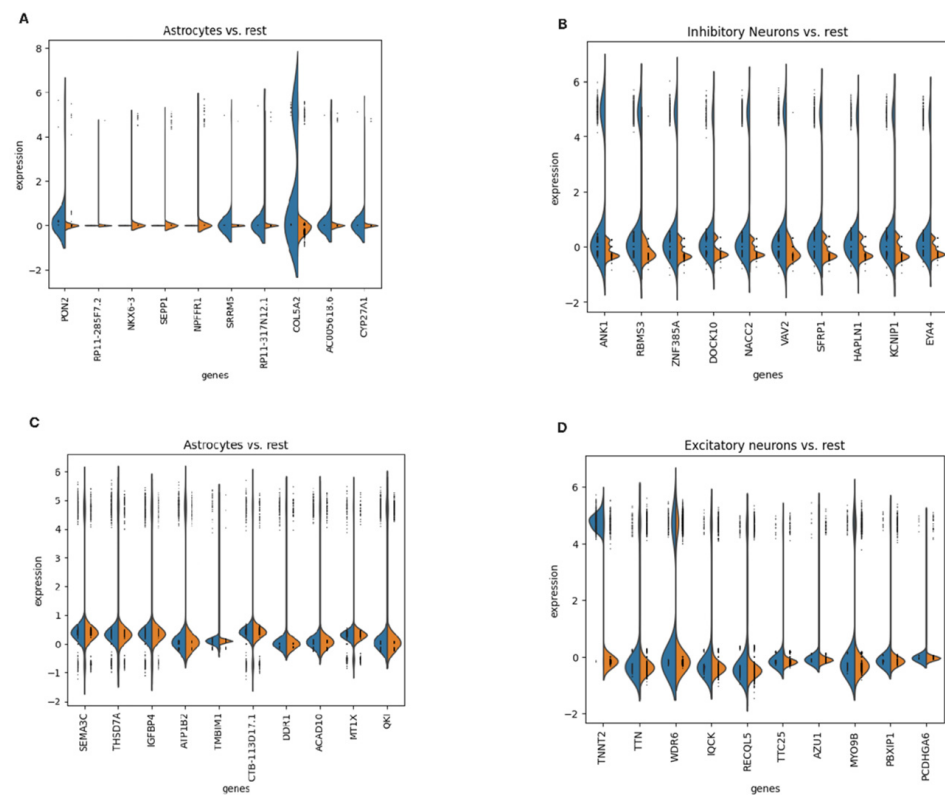


Figure 5. Violin plots display the distribution of expression levels for the top 10 differentially expressed genes (A) in cell-type astrocytes compared to the rest of the groups in the control dataset, (B) in cell-type inhibitor neurons compared to the rest of the groups in the control dataset, (C) in cell-type astrocytes compared to the rest of the groups in the disease dataset and (D) in excitatory neurons compared to the rest of the groups in the disease dataset.

Lastly, we merged the newly created AnnData (adata) objects for both the control and disease groups into a single adata object, following to the same analytical procedures as previously established (Figure 6). The integration facilitates a focused comparison of cells present in both the control and AD conditions, allowing us to pinpoint genes that exhibit significant differences between these groups. By concentrating on these particular cells, we tried to uncover crucial genetic markers that may elucidate the underlying mechanisms of disease progression or resilience. This comparative approach is instrumental in distinguishing the genetic expressions that are pivotal in disease manifestation compared to normal physiological states.

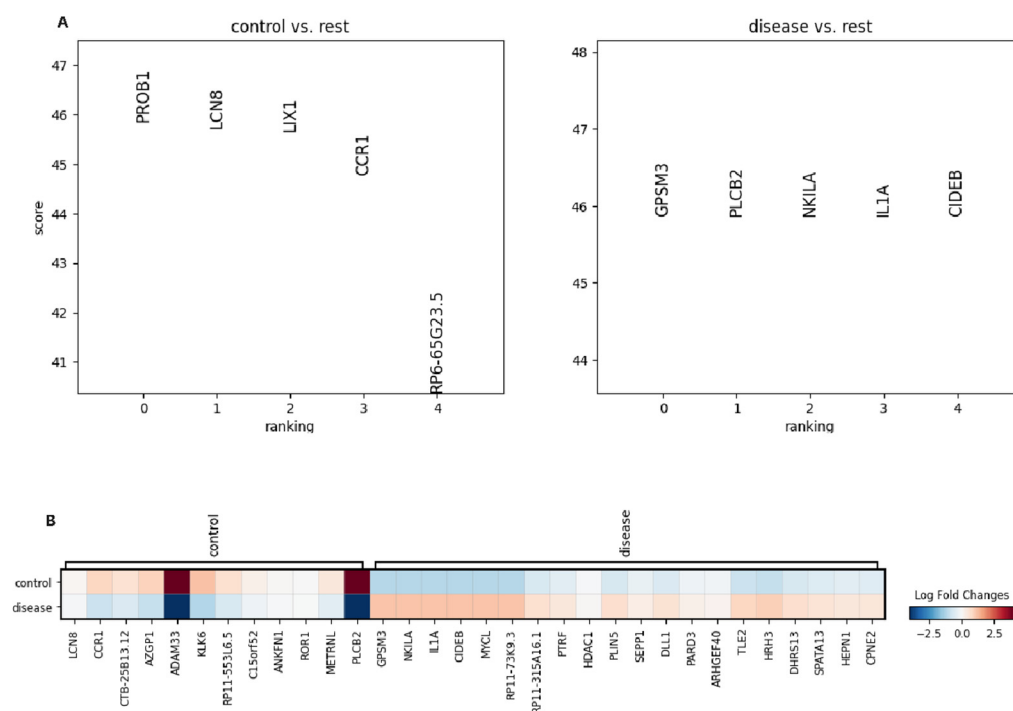


Figure 6. (A) Differential expression of the top five ranked genes in each condition compared to the other. (B) Heatmap of the log fold changes in gene expression between control and disease condition.

Across various classifiers, including K Neighbors Classifier, Extreme Gradient Boosting, Decision Tree Classifier, and Gradient Boosting Classifier, the analysis reveals consistently high values for accuracy and AUC scores. These metrics indicate robust discriminative capabilities in distinguishing between healthy and disease conditions based on the gene expression data (Figure 7). Furthermore, these classifiers demonstrate balanced performance in terms of recall, precision, and F1 score, highlighting their ability to effectively identify positive cases while minimizing false positives. Strong agreement metrics such as Kappa and MCC further validate the reliability of these classifiers, suggesting consistent and balanced predictions across different models. While there are variations in training times among the models, it's noteworthy that models like Extreme Gradient Boosting exhibit both high accuracy and efficiency, achieving remarkable results within a reasonable timeframe. Our analysis was facilitated by the PyCaret library (PyCaret version 3.3.2), a versatile and user-friendly tool for streamlined machine learning experimentation. PyCaret automates various aspects of the machine learning workflow, including data preprocessing, feature selection, model training, hyperparameter tuning, and model evaluation. Its intuitive interface and extensive suite of functionalities enable researchers to efficiently explore multiple models, compare their performance, and derive actionable insights from the data.

By leveraging PyCaret's capabilities, we were able to conduct a thorough evaluation of different classifiers on the gene expression dataset, facilitating informed decision-making and accelerating the research process. The library's comprehensive documentation and

extensive support for a wide range of machine learning tasks make it an invaluable resource for both novice and experienced practitioners in the field. Beyond the quantitative metrics, our computational analysis offers deeper insights into the underlying biological mechanisms.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
lightgbm	Light Gradient Boosting Machine	0.9854	0.9988	0.9854	0.9857	0.9854	0.9683	0.9686
knn	K Neighbors Classifier	0.9825	0.9958	0.9825	0.9832	0.9826	0.9623	0.9629
xgboost	Extreme Gradient Boosting	0.9810	0.9983	0.9810	0.9818	0.9811	0.9592	0.9598
dt	Decision Tree Classifier	0.9796	0.9779	0.9796	0.9796	0.9796	0.9557	0.9557
gbc	Gradient Boosting Classifier	0.9796	0.9981	0.9796	0.9804	0.9797	0.9561	0.9567
nb	Naive Bayes	0.9664	0.9959	0.9664	0.9698	0.9668	0.9290	0.9318
ridge	Ridge Classifier	0.9650	0.9732	0.9650	0.9657	0.9647	0.9230	0.9242
lda	Linear Discriminant Analysis	0.9650	0.9731	0.9650	0.9656	0.9648	0.9234	0.9243
qda	Quadratic Discriminant Analysis	0.9620	0.9935	0.9620	0.9650	0.9624	0.9194	0.9217
svm	SVM - Linear Kernel	0.9562	0.9746	0.9562	0.9574	0.9564	0.9059	0.9068
lr	Logistic Regression	0.9182	0.9783	0.9182	0.9259	0.9193	0.8285	0.8339
rf	Random Forest Classifier	0.9095	0.9997	0.9095	0.9280	0.9111	0.8141	0.8288
et	Extra Trees Classifier	0.8891	0.9999	0.8891	0.9158	0.8912	0.7750	0.7958
ada	Ada Boost Classifier	0.3606	0.9983	0.3606	0.1300	0.1911	0.0000	0.0000
dummy	Dummy Classifier	0.3606	0.5000	0.3606	0.1300	0.1911	0.0000	0.0000

Figure 7. A comprehensive analysis of performance metrics derived from various classifiers employing gene expression data from the condition dataset to classify individuals into healthy or disease conditions. Different trained models where the highlighted metrics indicate best score. The analysis was conducted using the PyCaret library.

In addition to these metrics, the confusion matrix offers a detailed view of each classifier's performance by illustrating the true positive, true negative, false positive, and false negative rates. Figure 8 shows the confusion matrix for the Custom Probability Threshold Classifier which highlights its impressive ability to correctly classify instances. This classifier refers to the method of converting predicted probabilities into class labels based on a specified threshold. By default, this threshold is set at 0.5, which means that any predicted probability above 0.5 is classified as positive, while those below 0.5 are classified as negative. In our analysis, we customized this threshold to 0.7 indicating that only predictions with a probability above 0.7 are classified as positive, making the classification criteria more stringent. The classifier achieved 978 true negatives indicating a high accuracy in identifying healthy samples. Furthermore, the 1726 true positives demonstrate its robustness in detecting disease conditions accurately. The minimal number of false negatives (28) and false positives (10) underscore the classifier's precision, ensuring that most positive identifications are correct while very few actual positive cases are missed. This high precision and recall combination corroborates the elevated F1 score observed in the evaluation metrics. The confusion matrix thus provides a granular insight into the classifier's performance, revealing its effectiveness in maintaining a low false-positive rate, which is crucial for reducing unnecessary follow-up tests and treatments. Simultaneously, the low false-negative rate ensures that most disease cases are correctly identified, highlighting the model's reliability and robustness in practical applications. These detailed breakdowns affirm the classifier's utility in real-world scenarios, where accurate and reliable predictions are paramount.

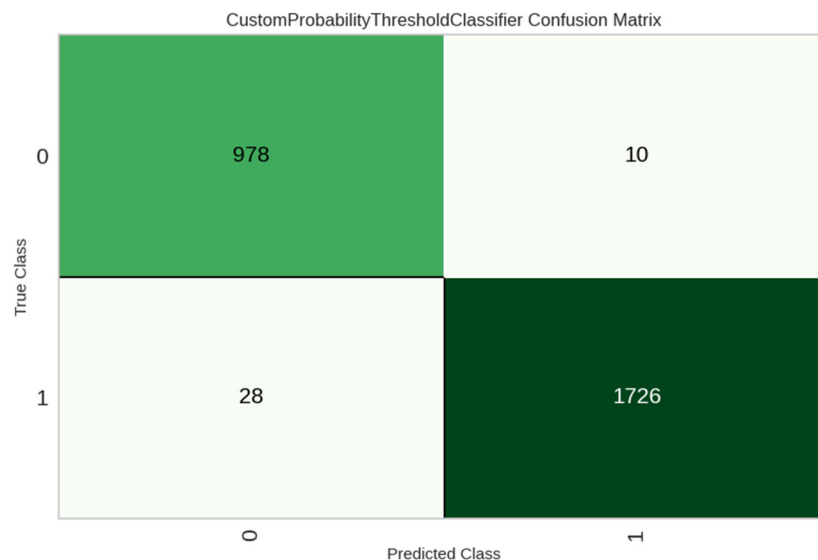


Figure 8. Confusion matrix for the Custom Probability Threshold Classifier.

We further utilized various libraries to identify pathways associated with upregulated and downregulated genes (Tables S1 and S2, respectively). Reactome pathway analysis showed that upregulated DEGs were mainly enriched in pathways such as G-alpha (I) signalling Events, transcription of neuronal ligands, interleukin-1 processing, caspase-mediated cleavage of cytoskeletal proteins and GPCR downstream signalling, while WikiPathway showed enrichment in neuroinflammation and glutamatergic signalling, interleukin-1 induced activation of NF-kB, and the IL-10 anti-inflammatory signalling pathway. The enriched gene ontology (GO) terms were divided into biological processes (BP), and the results of GO analysis revealed that upregulated DEGs were mainly enriched in BPs including positive regulation of gene expression, response to metal ions, and positive regulation of protein transport. On the other hand, downregulated genes were mainly enriched in neurotrophin, P75 NTR receptor-mediated signalling, LTC4-CYSLTR mediated IL4 production, and transcription of neuronal ligands, according to Reactome analysis and in Wnt signaling, the chemokine signaling pathway, and the leukotriene metabolic pathway (WikiPathway). BP analysis indicated that DEGs were significantly enriched in the leukocyte apoptotic process, glutathione catabolic process, and positive regulation of the ERK1 and ERK2 cascades.

This study provides a detailed analysis of AD by identifying key differentially expressed genes across multiple brain regions, shedding light on the intricate molecular dynamics associated with the disease's pathology. Our findings, particularly in the upregulation of genes like GSN and TTN in pathways associated with amyloidosis, muscle stretch and cardiac muscle contraction, reveal potent therapeutic targets. Additionally, the positive regulation of gene expression by genes such as CCDC88B and IL1A suggests new ways for modulating gene expression in positive regulation of T-cell maturation and inflammatory function. Notably, the identification of pathways related to neuroinflammation and glutamatergic signalling, featuring genes like IL1A, GRM4, and SST, emphasizes their potential role in AD's systemic pathological processes. These pathways are consistently altered across the studied regions, underscoring the importance of targeting these molecular mechanisms to mitigate the disease's progression. The regional expression of genes such as HDAC1 and ARHGEF40, involved in downregulation pathways like death receptor signalling and p75 NTR receptor-mediated signalling, hints at a complex regulatory mechanism that might confer specific regional vulnerabilities or resilience to AD pathology (Figure 9).

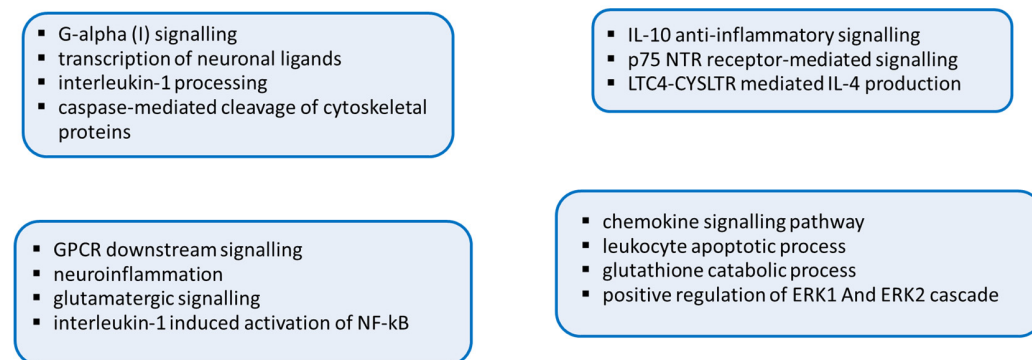


Figure 9. Pathways in which DEGs were primarily enriched according to our analysis.

4. Discussion

The present work includes data integration and analysis of specific scRNA-seq datasets from multiple brain regions associated with AD pathology. This approach seeks to provide a systemic view of the disease's impact across different human brain regions, leveraging data to uncover insights that have not been previously recognized due to the isolated nature of earlier analyses. This approach underscores the potential of computational analyses to deepen our understanding of AD from a holistic perspective, providing valuable insights that could lead to the development of targeted molecular interventions. Specifically, the study intends to achieve the following pillars: 1. **Data Integration:** Combine scRNA-seq datasets from critical brain regions such as the entorhinal cortex, prefrontal cortex, superior frontal gyrus, and superior parietal lobe. This integrated analysis will allow for a comparative assessment of cellular and molecular features across these regions, enhancing our understanding of how AD manifests differently in various parts of the brain. 2. **Computational Analysis:** Utilize computational methods to analyse these integrated datasets, focusing on identifying common and region-specific molecular signatures that characterize AD. This includes the application of batch effect correction, normalization, dimensionality reduction, and clustering algorithms to synthesize and interpret the complex data. 3. **Insight Development:** While not generating new experimental results, the study aims to derive novel insights into the pathology of AD by reanalysing existing data. This will include identifying patterns and correlations that may have been overlooked in previous studies that focused on single datasets or regions. 4. **Therapeutic Implications:** Explore potential therapeutic targets by understanding the molecular mechanisms across the brain's affected regions. Identifying pathways that are consistently altered in these regions could highlight targets for therapeutic interventions that might be effective across the broader spectrum of AD pathology. 5. **Methodological Contribution:** Demonstrate the power and utility of computational methods in the integration and analysis of complex and large-scale biological data. The study will showcase how computational approaches can be used to enhance the value of existing datasets, providing a blueprint for similar future studies in neurodegenerative diseases and beyond. By meeting these objectives, the study will significantly enrich our understanding of AD, offering value for future research into comprehensive and targeted treatments. It seeks to establish a new standard for the effective application of computational analysis in interpreting and integrating diverse biological data, thus paving the way for novel avenues in research and therapeutic advancements. According to Q-Q plot, significant deviations from the normal distribution are indicated, suggesting that gene expression data are not normally distributed. Additionally, the histogram shows a sharp peak around zero with a rapid drop-off, indicating that most data points are concentrated near this value, which further supports the non-normality of the distribution. Given these observations, it is clear that gene expression data do not follow a normal distribution. The Wilcoxon rank-sum test is a non-parametric method, meaning it does not assume normality in the data. Therefore, it is a suitable and robust choice for our analysis, allowing us to accurately identify differentially expressed genes.

without being affected by the non-normal distribution of the data. As Figure S4 shows, the significant deviations from normality observed in both the Q–Q plot and histogram justify the use of the Wilcoxon rank-sum test for differential gene expression analysis in our study.

While we refrain from making direct assertions about specific genes' diagnostic potential, our approach sheds light on the computational indicators that might point towards important genetic markers associated with the conditions under study. By leveraging advanced machine learning techniques and comprehensive evaluation strategies, our analysis provides a nuanced understanding of the gene expression patterns characteristic of different conditions. These findings not only contribute to our understanding of the molecular underpinnings of disease, but also offer valuable guidance for future research endeavours. Rather than as a conclusive diagnostic marker, our computational analysis serves as a powerful exploratory tool, indicating potential candidate genes worthy of further investigation. This nuanced approach underscores the importance of integrating computational methods with traditional experimental techniques in unravelling the complexities of disease mechanisms.

AD, the most prevalent cause of dementia among older adults, poses significant challenges due to its intricate and multifactorial nature. With genetic, environmental, and molecular factors contributing to it, unravelling the pathogenesis of AD and developing effective treatments is a persistent and complex endeavour [18,19]. The advent of scRNA-seq technology provides a methodology to explore the cellular heterogeneity of the tissue, by profiling tens of thousands of individual cells, and has opened new ways for exploring the molecular details of diseases with unprecedented precision [6]. More precisely, through scRNA-seq, researchers can probe the cellular diversity, offering a comprehensive approach to the specific cellular environmental conditions that contribute to disease progression [20]. Recent technological advances have particularly enhanced our ability to discern subtle variations in gene expression across individual cells, which is crucial for identifying the molecular signatures associated with AD. However, the use of scRNA-seq in AD has been primarily restricted to isolated analyses of specific brain regions or datasets [21,22]. A comprehensive, integrated examination across multiple affected regions remains rare, which limits our understanding of the systemic and regional impacts of the disease across the brain's complex landscape.

By integrating data, computational analysis revealed novel molecular signatures, validating observed patterns as authentic biological phenomena rather than artifacts of data manipulation. This step is necessary in preventing any overlap or confusion during the merging process, guaranteeing that each cell, now part of a larger dataset, retains a distinct identity. This clear delineation is fundamental for subsequent analyses, ensuring that data from disparate datasets can be accurately compared. This rigorous approach not only bolsters the credibility of our findings but also establishes a methodological blueprint for future studies aiming to decode the complex molecular landscape of AD. By integrating insights from external studies indicating distinct transcriptional networks in AD, particularly within neuronal and glial populations, we corroborate our findings within a broader scientific context [10]. Furthermore, the dynamic perspective on gene expression, highlighted through RNA velocity studies, complements our static analysis by illustrating the importance of temporal dynamics in understanding cellular responses in AD [14].

Blood-based biomarkers, especially immune-related ones, could provide a more accessible and cost-effective solution for early AD detection. In a recent work, advances in understanding brain–immune interactions and how machine learning can combine various biomarkers and demographic information to improve early diagnosis is discussed. Furthermore, mechanistic modelling techniques for analysing cell dynamics are explored, highlighting the potential of immune-related blood biomarkers for early AD diagnosis [23]. The clinical implications of discovering new diagnostic markers or therapeutic targets are crucial. Baheti et al. highlight the advantages of molecular modelling methods, which offer a faster and more efficient way to design drugs with improved efficacy and ethical considerations compared to traditional approaches. Researchers are increasingly adopting these advanced methods to better address AD and other diseases [24]. Looking forward,

the framework developed in the present work promises to be a robust analytical tool for comparing cellular and molecular changes between AD patients and healthy controls. This comparative analysis can shed light not only the specific pathological triggers associated with AD but also on potential resilience factors found within the control group. Such insights could inspire the development of focused interventions aimed at replicating these resilience factors in susceptible populations. Moreover, by harnessing this methodology, future research can leverage scRNA-seq data to gain a systemic view of AD's impact across different brain regions [25]. This approach will enable a deeper understanding of the disease at a cellular level, paving the way for precision medicine strategies that are fine-tuned to the molecular profiles observed in individual patients [26]. However, this study does face limitations, primarily due to its reliance on existing datasets, which may not capture the full spectrum of cellular diversity in AD pathology. The analytical methods, while sophisticated, also depend heavily on the quality and completeness of the data integrated into our study. Future research should aim to include more diverse datasets, potentially incorporating longitudinal data to observe the progression of AD over time, which could provide further insights into the dynamics of the disease's development.

5. Conclusions

Herein, a comprehensive approach for integrating scRNA-seq datasets from various control samples corresponding to different regions of the human brain was carried out. The analysis was performed in order to identify commonalities or discrepancies across distinct brain regions, which requires meticulous preprocessing to ensure data quality and comparability. Through this process, the capabilities of scRNA-seq to bridge this gap were harnessed by providing a detailed, cross-regional analysis of the brain regions most affected by AD, including the entorhinal cortex, prefrontal cortex, superior frontal gyrus, and superior parietal lobe. By utilizing sophisticated machine learning methods, conducting a thorough evaluation of different classifiers on the gene expression dataset and thorough evaluation approaches, our analysis offers detailed insight into the gene expression patterns specific to various conditions. The present computational analysis, empowered by the PyCaret library, offers valuable insights into the genetic signatures associated with different conditions and enables us to assess the cellular and molecular alterations between individuals with AD and healthy control subjects, highlighting the unique and overlapping pathways of degeneration across different brain regions. By merging and analysing data from disparate regions, the present study leverages existing scRNA-seq datasets highlighting the potential of computational analysis to provide deep insights into the complex biology of AD from a multi-regional point of view. Moreover, an integration of control and diseased datasets of all regions was performed, and gene expression disparities between healthy cells and those of the AD-affected condition were conducted. These findings pave the way for further research into identifying and validating key genetic markers for diseases, ultimately advancing our understanding of disease pathology and informing future clinical diagnostics and therapeutic interventions.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/info15090523/s1>; Figure S1: Differential expression of the top 10 ranked genes in each identified cell group compared to the rest, as analysed in the control dataset; Figure S2: Dot plots show the cells category based on most expressed genes in the control group; Figure S3: Dot plots show the cells category based on most expressed genes in the disease group; Figure S4: Q-Q plot and histogram justify the use of the Wilcoxon rank-sum test for differential gene expression analysis in our study; Table S1: Pathways for upregulated genes; Table S2: Pathways for downregulated genes.

Author Contributions: Conceptualization, M.G.K.; methodology, P.K. and M.G.K.; software, P.K.; validation, P.K. and M.G.K.; formal analysis, P.K. and M.G.K.; data curation, P.K. and M.G.K.; writing—original draft preparation, P.K. and M.G.K.; writing—review and editing, T.P.E. and P.V.; funding acquisition, P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union-Next Generation EU, Greece 2.0 National Recovery and Resilience Plan Flagship program TAEDR-0535850.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used already available data from the scREAD database. No new data were created.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fan, L.; Mao, C.; Hu, X.; Zhang, S.; Yang, Z.; Hu, Z.; Sun, H.; Fan, Y.; Dong, Y.; Yang, J.; et al. New insights into the pathogenesis of Alzheimer's disease. *Front. Neurol.* **2020**, *10*, 1312. [\[CrossRef\]](#)
2. Guo, T.; Zhang, D.; Zen, Y.; Huang, T.Y.; Xu, H.; Zhao, Y. Molecular and cellular mechanisms underlying the pathogenesis of Alzheimer's disease. *Mol. Neurodegener.* **2020**, *15*, 1–37.
3. Breijyeh, Z.; Karaman, R. Comprehensive review on Alzheimer's disease: Causes and treatment. *Molecules* **2020**, *25*, 5789. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Chen, G.; Ning, B.; Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **2019**, *10*, 441123. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Angerer, P.; Simon, L.; Tritschler, S.; Wolf, F.A.; Fischer, D.; Theis, F.J. Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **2017**, *4*, 85–91. [\[CrossRef\]](#)
6. Jovic, D.; Liang, X.; Zeng, H.; Lin, L.; Xu, F.; Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.* **2022**, *12*, e694. [\[CrossRef\]](#)
7. Luquez, T.; Gaur, P.; Kosater, I.M.; Lam, M.; Lee, D.I.; Mares, J.; Paryani, F.; Yadav, A.; Menon, V. Cell type-specific changes identified by single-cell transcriptomics in Alzheimer's disease. *Genome Med.* **2022**, *14*, 136. [\[CrossRef\]](#)
8. Saura, C.A.; Deprada, A.; Capilla-López, M.D.; Parra-Damas, A. April. Revealing cell vulnerability in Alzheimer's disease by single-cell transcriptomics. *Semin. Cell Dev. Biol.* **2023**, *139*, 73–83. [\[CrossRef\]](#)
9. Lampinen, R.; Fazaludeen, M.F.; Avesani, S.; Örd, T.; Penttilä, E.; Lehtola, J.M.; Saari, T.; Hannonen, S.; Saveleva, L.; Kaartinen, E.; et al. Single-cell RNA-Seq analysis of olfactory mucosal cells of Alzheimer's disease patients. *Cells* **2022**, *11*, 676. [\[CrossRef\]](#)
10. Soreq, L.; Bird, H.; Mohamed, W.; Hardy, J. Single-cell RNA sequencing analysis of human Alzheimer's disease brain samples reveals neuronal and glial specific cells differential expression. *PLoS ONE* **2023**, *18*, e0277630. [\[CrossRef\]](#)
11. Aslanis, I.; Krokidis, M.G.; Dimitrakopoulos, G.N.; Vrahatis, A.G. Identifying Network Biomarkers for Alzheimer's Disease Using Single-Cell RNA Sequencing Data. In *Worldwide Congress on "Genetics, Geriatrics and Neurodegenerative Diseases Research"*; Springer International Publishing: Cham, Switzerland, 2022; pp. 207–214.
12. Krokidis, M.G.; Vrahatis, A.G.; Lazaros, K.; Vlamos, P. Exploring Promising Biomarkers for Alzheimer's Disease through the Computational Analysis of Peripheral Blood Single-Cell RNA Sequencing Data. *Appl. Sci.* **2023**, *13*, 5553. [\[CrossRef\]](#)
13. Pushparaj, P.N.; Kalamegam, G.; Wali Sait, K.H.; Rasool, M. Decoding the role of astrocytes in the entorhinal cortex in Alzheimer's disease using high-dimensional single-nucleus RNA sequencing data and next-generation knowledge discovery methodologies: Focus on drugs and natural product remedies for dementia. *Front. Pharmacol.* **2022**, *12*, 720170. [\[CrossRef\]](#)
14. Adewale, Q.; Khan, A.F.; Bennett, D.A.; Iturria-Medina, Y. Single-nucleus RNA velocity reveals critical synaptic and cell-cycle dysregulations in neuropathologically confirmed Alzheimer's disease. *Sci. Rep.* **2024**, *14*, 7269. [\[CrossRef\]](#)
15. Guennewig, B.; Lim, J.; Marshall, L.; McCorkindale, A.N.; Paasila, P.J.; Patrick, E.; Kril, J.J.; Halliday, G.M.; Cooper, A.A.; Sutherland, G.T. Defining early changes in Alzheimer's disease from RNA sequencing of brain regions differentially affected by pathology. *Sci. Rep.* **2021**, *11*, 4865.
16. Jiang, J.; Wang, C.; Qi, R.; Fu, H.; Ma, Q. scREAD: A single-cell RNA-Seq database for Alzheimer's disease. *iScience* **2020**, *23*, 101769. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Krokidis, M.G.; Vrahatis, A.G.; Lazaros, K.; Skolariki, K.; Exarchos, T.P.; Vlamos, P. Machine Learning Analysis of Alzheimer's Disease Single-Cell RNA-Sequencing Data across Cortex and Hippocampus Regions. *Curr. Issues Mol. Biol.* **2023**, *45*, 8652–8669. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Ibanez, L.; Cruchaga, C.; Fernández, M.V. Advances in genetic and molecular understanding of Alzheimer's disease. *Genes* **2021**, *12*, 1247. [\[CrossRef\]](#)
19. Zhang, X.X.; Tian, Y.; Wang, Z.T.; Ma, Y.H.; Tan, L.; Yu, J.T. The epidemiology of Alzheimer's disease modifiable risk factors and prevention. *J. Prev. Alzheimer's Dis.* **2021**, *8*, 313–321. [\[CrossRef\]](#)
20. Choi, Y.H.; Kim, J.K. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol. Cells* **2019**, *42*, 189–199.
21. Piwecka, M.; Rajewsky, N.; Rybak-Wolf, A. Single-cell and spatial transcriptomics: Deciphering brain complexity in health and disease. *Nat. Rev. Neurol.* **2023**, *19*, 346–362. [\[CrossRef\]](#)
22. Mathys, H.; Davila-Velderrain, J.; Peng, Z.; Gao, F.; Mohammadi, S.; Young, J.Z.; Menon, M.; He, L.; Abdurrob, F.; Jiang, X.; et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **2019**, *570*, 332–337. [\[CrossRef\]](#) [\[PubMed\]](#)

23. Krix, S.; Wilczynski, E.; Falgàs, N.; Sánchez-Valle, R.; Yoles, E.; Nevo, U.; Baruch, K.; Fröhlich, H. Towards early diagnosis of Alzheimer's disease: Advances in immune-related blood biomarkers and computational approaches. *Front. Immunol.* **2024**, *15*, 1343900. [[CrossRef](#)] [[PubMed](#)]
24. Baheti, K.; Kale, M. Methodologies Related to Computational Models in View of Developing Anti-Alzheimer Drugs: An Overview. *Curr. Drug Discov. Technol.* **2019**, *16*, 66–73. [[CrossRef](#)] [[PubMed](#)]
25. Johnson, T.S.; Xiang, S.; Helm, B.R.; Abrams, Z.B.; Neidecker, P.; Machiraju, R.; Zhang, Y.; Huang, K.; Zhang, J. Spatial cell type composition in normal and Alzheimer's human brains is revealed using integrated mouse and human single cell RNA sequencing. *Sci. Rep.* **2020**, *10*, 18014. [[CrossRef](#)]
26. Kim, D.; Tran, A.; Kim, H.J.; Lin, Y.; Yang, J.Y.H.; Yang, P. Gene regulatory network reconstruction: Harnessing the power of single-cell multi-omic data. *NPJ Syst. Biol. Appl.* **2023**, *9*, 51. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.