Cohort-level differential distributional analysis for studying microglia in Alzheimer's disease via single-nuclei RNA-sequencing

Wenjing Tati Zhang[1], **Katherine E. Prater**[2], **and Kevin Z. Lin**[1,*]

[1]University of Washington, Biostatistics, Seattle, 98195, USA
[2]University of Washington, Neurology, Seattle, 98195, USA
[*]kzlin@uw.edu

ABSTRACT

Microglia, the brain's resident immune cells, play a critical role in Alzheimer's disease (AD) by mediating neuroinflammation and cellular dysfunction. Understanding their transcriptomic diversity is vital for identifying new therapeutic targets. Single-nuclei RNA sequencing enables a detailed study of microglia in Alzheimer's disease (AD), but current analytical approaches face two key limitations: uncertainty about the consistency of differential expression methods across varying cohort sizes and a narrow focus on mean expression differences that may miss other biologically relevant patterns. Here, we present Was2CoDE (Wasserstein-2 based Cohort Differential Expression), a framework that comprehensively analyzes distributional differences in gene expression between AD and control donors. By comparing DE methods (NEBULA, eSVD-DE, and DESeq2 with pseudobulking) across three microglia datasets from the prefrontal cortex, ranging from 22 to 345 donors, we demonstrate that conventional differential expression methods show increasing agreement as cohort size grows. We further identify that differences between AD and control donors extend beyond mean expression changes, with significant variance heterogeneity in immune-related pathways suggesting dysregulated microglia responses. Was2CoDE's mathematical decomposition of the Wasserstein-2 distance reveals genes and pathways not detected by standard differential expression methods while maintaining correlations with established approaches. Our results provide a unified framework for analyzing cohort-level single-nucleus RNA sequencing data and reveal new insights into microglia heterogeneity in AD, with implications for therapeutic development targeting neuroinflammation.

Introduction

Microglia are increasingly recognized as pivotal players in the progression of Alzheimer's disease (AD), positioning them as promising therapeutic targets[1]. In AD, microglia release inflammatory mediators that disrupt neuronal and glial function, lose their neuroprotective roles, and aberrantly phagocytose synapses and neurons. Additionally, experimental models suggest microglia facilitate tau protein spread and serve as primary mediators of amyloid-beta ($A\beta$) clearance, including reductions achieved through antibody-based immunotherapies. Recent studies have revealed microglia heterogeneity, with sub-types expressing homeostatic and activated markers[1]. However, microglia's diverse and intricate functions still need to be better understood, complicating efforts to develop targeted therapeutics. Understanding the transcriptomic phenotypes and gene regulatory networks underlying these dynamic microglia states is crucial for advancing therapeutic strategies to address neuroinflammation in AD.

One major challenge in studying microglia in AD is obtaining detailed transcriptomic data from human brains. Traditional bulk RNA sequencing methods of brain tissue cannot detect microglia-specific transcriptional differences between AD and non-AD donors since microglia are much smaller in size than neurons and also represent a much smaller percentage of cell types compared to other brain cell types. Compounding this difficulty, factors such as post-mortem intervals and individual variability further complicate analyses. Single-nucleus RNA sequencing has emerged as a powerful tool to enrich microglia nuclei from AD and control brain samples, enabling a deeper examination of microglia molecular phenotypes. This approach has provided valuable insights into the gene networks and regulatory pathways driving these transcriptomic changes. Investigating differential expression within microglia offers a promising avenue to uncover pathways and regulatory networks critical to AD pathology, paving the way for more precise therapeutic interventions. However, there remain two major limitations of existing single-nuclei workflows, both of which constitute the main focus of this paper.

First, although numerous differential expression (DE) methods have been developed, each method is grounded in distinct statistical and computational frameworks[2]. There is a lack of consensus persists regarding whether these methods yield consistent biological insights, particularly as the number of donors in the cohort expands. A critical distinction among these methods lies in how they adjust for donor-specific covariates when testing for phenotypic effects. For example, NEBULA employs a negative binomial mixed model to account for donor covariates[3], while eSVD-DE uses matrix factorization to pool information across genes and remove confounding effects[4]. Alternatively, bulk RNA sequencing methods such as DESeq2[5] can be adapted to single-nuclei data through pseudo-bulking, where nuclei from each donor are aggregated[6]. Although larger cohorts theoretically improve the accuracy of these methods, this statistical intuition remains underexplored. Moreover, the scientific challenge of human variability translates into statistical complexity, as factors such as age, sex, and post-mortem interval must be carefully adjusted in sparse single-nuclei data—a nontrivial task when variability is high across donors.

Second, another limitation of existing differential expression methods is their focus on differential mean expression, which may overlook other biologically relevant patterns. For instance, differences in gene expression variance may reflect underlying dynamics of cellular coordination or variation in cell-state composition among donors. Such patterns could illuminate protective mechanisms or unique phenotypic traits in specific donor subsets. While the IDEAS framework[7] represents a step forward by considering broader distributional differences, it remains unclear how its approach integrates with methods focused on differential mean expression. There is a pressing need for a comprehensive framework synthesizing findings from diverse differential expression methods to provide more nuanced biological insights. This unification is particularly vital for microglia, where understanding the transcriptomic diversity could unlock new therapeutic pathways for AD.

Results

Single-nuclei RNA-seq Microglia Data Acquisition

We first describe the different datasets of microglia that we investigate in this paper, each of single-nuclei RNA-sequencing data of frozen post-mortem tissue from the prefrontal cortex (PFC) brain region but have different experimental characteristics. This allows us to better compare our differential expression results across datasets while investigating how the number of donors or nuclei influences different statistical methods. The microglia samples for this study were obtained from three datasets: Prater et al. (2023)[1], the SEA-AD cohort[8], and ROSMAP cohort[9]. Each dataset comprises a cohort of donors diagnosed with Alzheimer's disease based on the donor's neuropathology (i.e., "cases") and a control group of cognitively normal individuals (i.e., "controls").

Figure 1 highlights the key summary statistics of each dataset, and we highlight some key distinctions across the datasets:

- The dataset from Prater et al.[1] consists of 22 donors (12 cases, 10 controls). The donors were recruited through the Neuropathology Core of the Alzheimer's Disease Research Center at the University of Washington. This study used fluorescence-activated nuclei sorting (FANS) using the myeloid-specific transcription factor PU.1 to enrich microglia nuclei from flash-frozen post-mortem tissue from the PFC (specifically the dorsolateral prefrontal cortex). This yielded a substantial enrichment of microglia nuclei, verified by markers like CX3CR1, C1QB, and SPI1. This is demonstrated in Figure 1, where we see a substantially higher number of microglia per donor than the other datasets in this paper. We refer to this dataset as the "Prater dataset."

- The SEA-AD (Seattle Alzheimer's Disease) dataset[8] consists of 76 donors (57 cases, 21 controls). The donors were recruited via the Adult Changes in Thought (ACT) study, a community cohort study of older adults from Kaiser Permanente Washington, and the University of Washington Alzheimer's Disease Research Center (ADRC). This study also used FANS to isolate nuclei, with a focus on a 70% neuron to 30% non-neuronal nuclei ratio from fresh frozen tissue of the PFC (specifically, the dorsal frontal cortex) to better capture non-neuronal populations like microglia. The nuclei, including the microglia, in this paper, were labeled via scANVI[10], a label-transfer method based on a neurotypical reference brain from the BRAIN Initiative Cell Census Network (BICCN). Compared to the other datasets in this paper, this study has a high number of microglia per donor, given the modestly large number of donors.

- The ROSMAP dataset[9] consists of 345 donors from the ROSMAP study[11] (192 cases, 153 control). This study combines two community-based studies: the Religious Orders Study (ROS), which recruits nuns, priests, and brothers from across the United States, and the Rush Memory and Aging Project (MAP), which targets the general population primarily from northeastern Illinois. This study sequenced the fresh frozen post-mortem tissue from the PFC, and the microglia were labeled based on the gene expression of immunological markers such as CSF1R, CD74, and C3. Compared to the other datasets in this paper, this study has the highest number of donors and average sequencing depth per microglia.

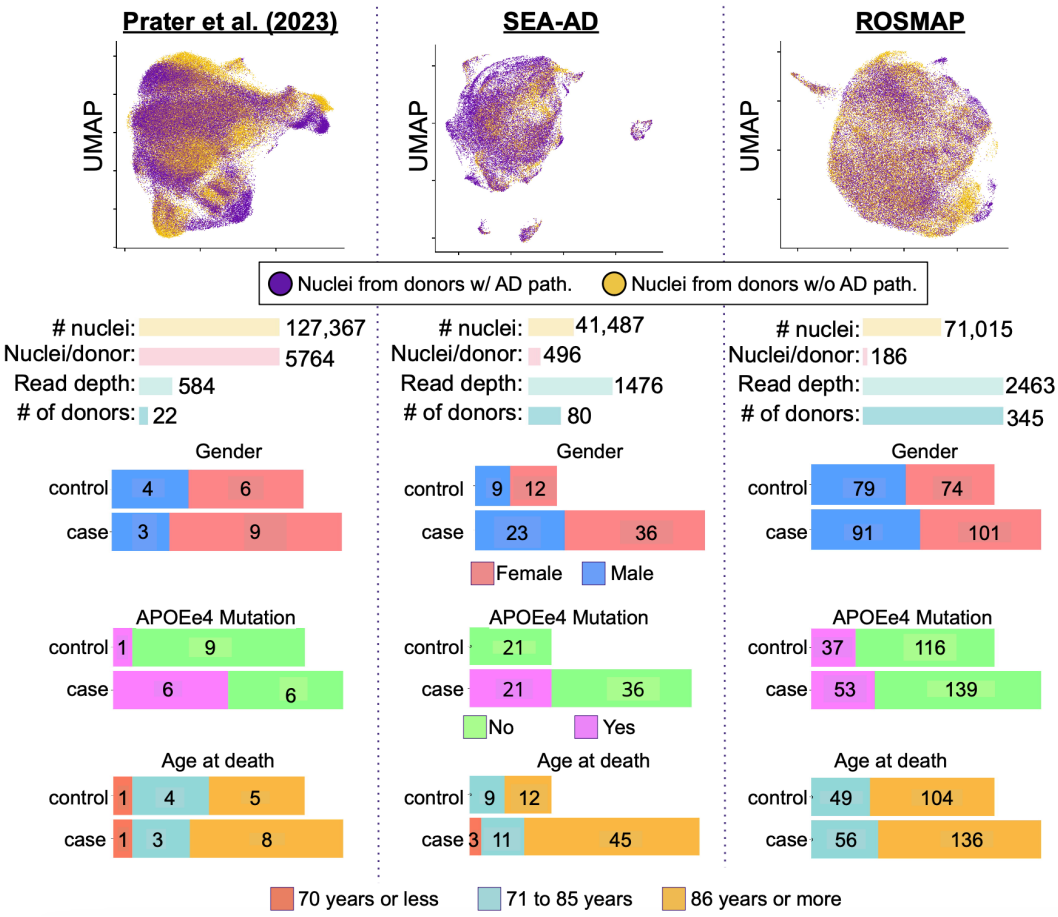

Figure 1. Comparison of three single-nuclei RNA sequencing datasets used in Alzheimer's disease (AD) research: Prater et al. (2023)[1], SEA-AD[8], and ROSMAP[9]. UMAP visualizations (top) display the transcriptional profiles from the three datasets, with nuclei colored by donor AD pathology status (purple: donors with AD pathology; yellow: donors without AD pathology). Dataset characteristics are shown (middle), including total nuclei count, mean nuclei per donor, sequencing read depth, and number of donors. Demographic distributions are shown (bottom), stratified by case-control status, showing gender composition (male/female), APOE4 mutation carrier status (Yes/No), and age at death ($\leq$70, 71-85, $\geq$86 years) for each cohort. The datasets demonstrate varying scale and donor characteristics: Prater et al. (2023)[1] features deep cellular profiling (5,764 nuclei/donor) from 22 donors; SEA-AD comprises moderate coverage (496 nuclei/donor) across 80 donors; and ROSMAP provides broad donor sampling (345 donors) with targeted cellular profiling (186 nuclei/donor).

Investigating the relationship between number of donors and agreement between methods

Equipped with these datasets, we start our investigation on the first primary analyses in this paper – as cohorts recruit more donors, do different statistical methods result in more similar findings? This hypothesis is rooted in statistical theory, where many mathematical analyses across the statistics field show that the distinction between different estimators is driven by *statistical power*[12–14]. That is, certain methods can find meaningful patterns before others when used on the same dataset thanks to clever statistical "tricks" to better use the data. Therefore, we wonder if this premise is similar for differential expression from cohort-level analyses. In particular, if the distinction between NEBULA and eSVD-DE is primarily an issue of statistical power, then, conceptually, if the cohort size were large enough, then the choice of which method is used would be moot since any sensible method would find similar findings. If true, our finding would dramatically enhance existing cohort-wide scRNA-seq analysis workflows – this would demonstrate that recruitment of donors would have a much more substantial impact on the biological findings than the choice of computational workflow.

We set out to investigate this hypothesis by performing DESeq2 with pseudobulking, NEBULA, and eSVD-DE on each dataset, and we observe that the findings across different methods generally become more correlated as the number of donors increases. Figure 2 shows our results, where each scatter plot displays the pairwise comparisons of the signed negative log-10 p-values between any two methods analyze the three datasets: the Prater dataset (small number of donors), SEA-AD (medium number), and ROSMAP (large number). The "sign" for these p-values is based on the sign of the estimated log

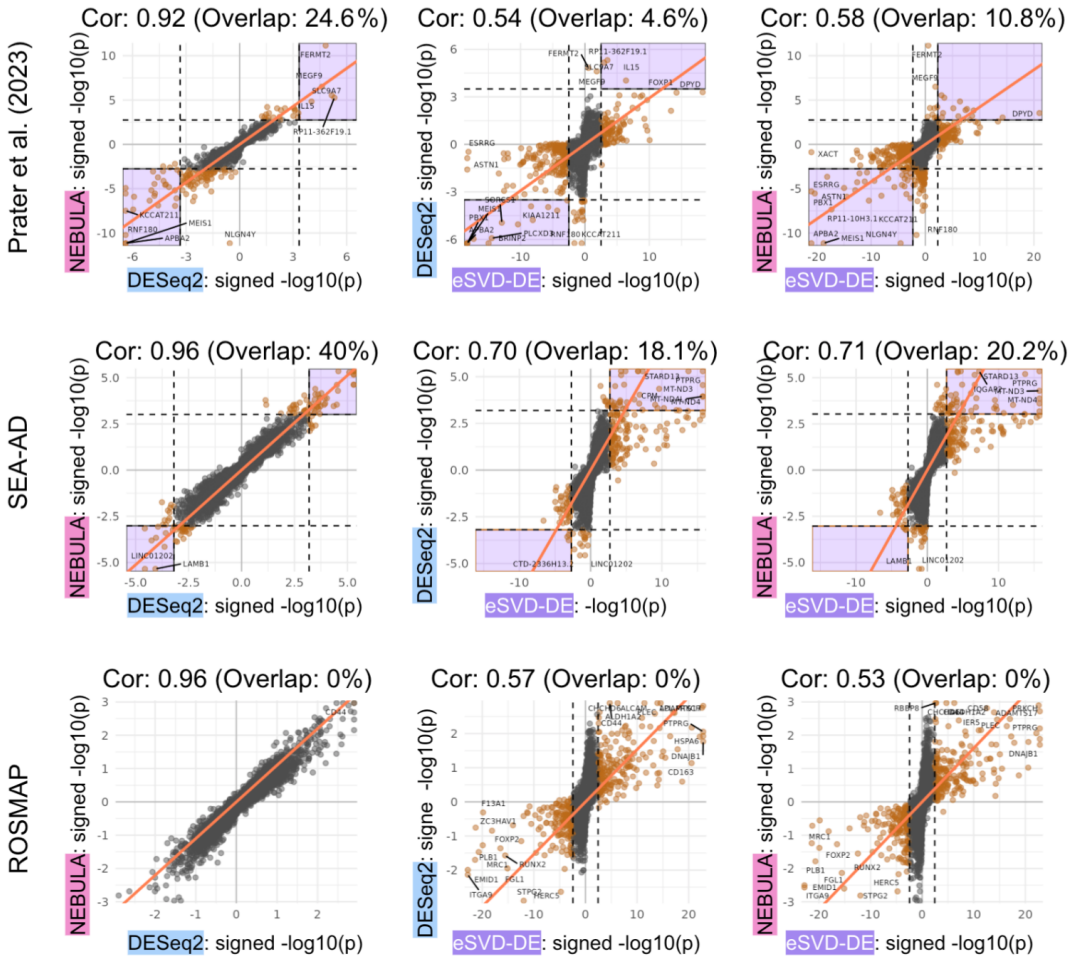

Figure 2. Correlation analysis comparing differential expression results across three methods: NEBULA, DESeq2, and eSVD-DE. Results are shown for three datasets: Prater, SEA-AD, and ROSMAP. Each scatter plot displays signed -log10(p-values) between pairs of methods, with correlation coefficients indicated at the top of each panel. NEBULA shows consistently high correlation with DESeq2 (Cor: 0.92-0.96), while correlations between other method pairs are more moderate (Cor: 0.53-0.71), suggesting strong agreement between NEBULA and DESeq2 methodologies across different datasets. The solid orange line denotes the $x = y$ line for visual reference. We report the ratio between the number of significant genes detected by *both* methods over the number of significant genes detected by *either* method ("Overlap").

fold change (LFC) based on that method. The more a gene's signed negative log-10 p-value deviates from 0 (i.e., has a large magnitude), the more significant that gene is deemed to be differentially expressed between cases and controls. Our results generally demonstrate a trend that more donors result in a higher correlation in p-values. We make two initial observations. First, many methods agree on the directionality of the LFC. Second, the correlation between DESeq2 with pseudobulking and NEBULA is quite high, often deeming many of the same genes significantly differentially expressed between cases and controls. Finally, we note that the correlation between the methods' estimated LFC is often higher than the correlation between the method's signed negative log-10 p-values.

The comparison of differential expression analysis methods reveals a high correlation between NEBULA and DESeq2 in their results across multiple datasets, suggesting that both methods consistently capture overlapping gene expression patterns. However, although the ROSMAP cohort has the highest number of donors, there is less agreement between methods when compared to the SEA-AD cohort. Our findings suggest three potential hypotheses to explain this phenomenon:

1. **Potential importance of nuclei per donor**: Based on Figure 1, the average nuclei per donor in the SEA-AD cohort is more than twice as high compared to the ROSMAP cohort. A higher number of nuclei per donor means the statistical method has more certainty in measuring the mean expression of each donor.

2. **Potential importance of microglia sub-types**: Based on Figure 1, we see evidence that there are multiple microglia sub-types. This is reported in multiple papers[1,9]. Our analysis is on all the microglia simultaneously, but different

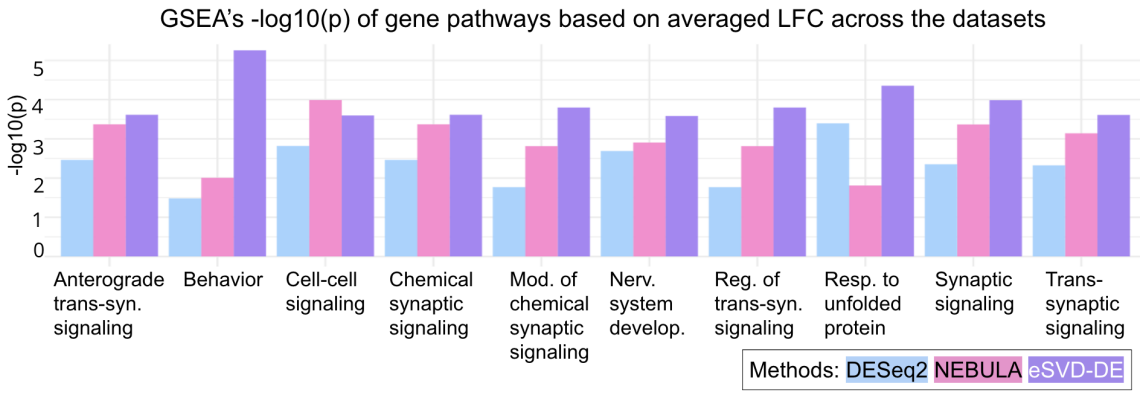

Figure 3. Gene Set Enrichment Analysis (GSEA) of log fold-change values averaged across the three datasets, comparing the three methods: DESeq2, NEBULA, and eSVD-DE. The analysis shows enrichment scores ($-\log_{10}(p)$) for ten neural signaling pathways. The bar colors the pathways from each analysis. The shown pathways are representative of the overall finding that eSVD-DE yielded many more significant pathways (with multiple-testing correction) compared to DESeq2 and NEUBLA.

sub-type compositions in each cohort could confound the biological differences between case and control donors.

3. **Potential importance of statistical target**: Since DESeq2 with pseudobulking, NEBULA, and eSVD-DE all use different statistical models and estimators, it is possible that we should not naturally expect methods to yield correlated results, even as cohort sizes increase.

In future work, we plan to tease apart these three hypotheses to explain our findings in Figure 2.

Investigating shared biological findings across all three datasets

Another angle to investigate the reliability of a statistical method is to assess if a method finds biologically meaningful pathways *across* datasets. Since the three datasets are all of microglia from the PFC brain region based on frozen brain tissue, comparing the findings across datasets is biologically appropriate.

To investigate this, we performed Gene Set Enrichment Analysis (GSEA) to evaluate the performance of three differential expression analysis methods: DESeq2, NEBULA, and eSVD-DE. Our results demonstrate that eSVD-DE consistently exhibits more sensitivity to certain pathways, achieving the highest enrichment scores across most pathways examined. These pathways generally were related to cell signaling and neurobiological pathways (Fig. 3). Overall, the concordance in the relative importance of pathways across different methods implies that, while sensitivity levels vary, the overall patterns of pathway enrichment remain similar. NEBULA generally produces enrichment scores with moderate significance, serving as a middle ground between the more sensitive eSVD-DE and the more conservative DESeq2, yielding the lowest scores among the methods assessed. Despite these variations in sensitivity, all three methods show robust enrichment in the cell-cell signaling pathway, suggesting a strong and consistent biological signal that is detectable regardless of the analytical approach employed. For example, the synaptic signaling-related pathways exhibit consistent moderate to strong enrichment across all methods, suggesting the reliability of detecting these critical neurobiological processes.

There are significant differences in the variance, beyond differences in the mean

While all of our analyses above focused on *differential mean* in gene expression (i.e., a statistical difference in the average gene expression between case and control donors), we also explore if there are *differential variance* in gene expression (i.e., a statistical difference in the variance of gene expression between case and control donors). This investigation stems from recent scRNA-seq analyses that found that gene coordination decreases as cells age[15–18]. Regarding AD, we hypothesize that donors with AD pathology have microglia that "age faster" or undergo accelerated disorganization. This can statistically manifest as donors with AD pathology have a higher variance in certain genes than donors without AD pathology.

To investigate differences in gene expression variance between AD patients and control donors, we compare the variability of gene expression levels for each gene between case and control donors. The volcano plot of the SEA-AD dataset reveals substantial gene expression variability, with several genes showing significant differences in variance (Fig. 4A). Among the significantly variable genes, notably, PTPRG (protein tyrosine phosphatase receptor type G) and CELSR1 (cadherin EGF LAG seven-pass G-type receptor 1) exhibit markedly increased variance in the disease state and have been implicated with AD[19,20]. Most genes with significant variance changes show positive log fold changes, indicating increased expression

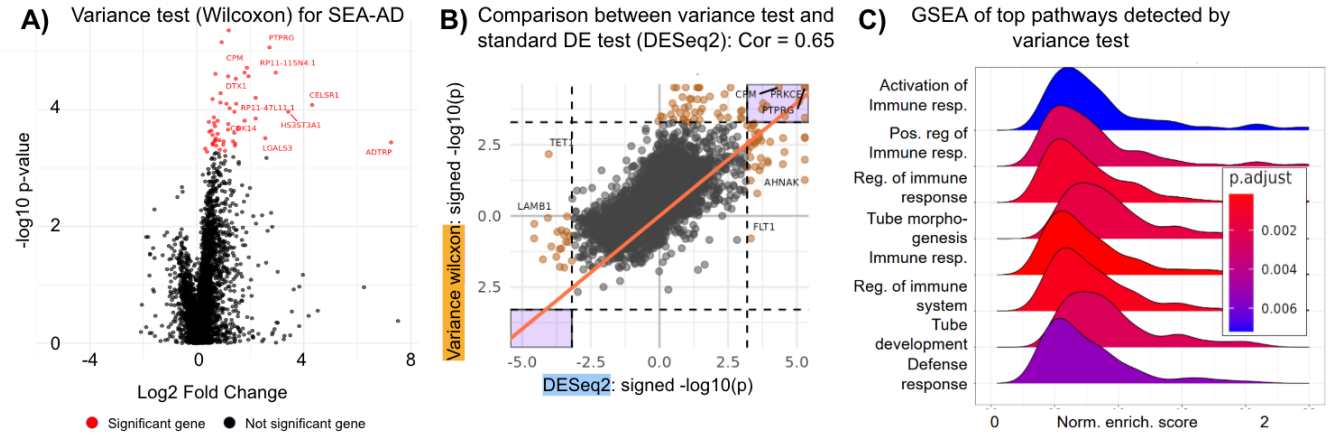

Figure 4. **A**) Volcano plot where the statistically significant genes are marked in red. **B**) Correlation between the signed negative $\log_{10}$ p-values between the differential variance test and DESeq2. **C**) GSEA of the top pathways based on the differential variance analysis shows each pathway's normalized enrichment scores. The color intensity indicates the significance of enrichment (adjusted p-value). All plots are based on the SEA-AD dataset.

variability in the disease condition. Although the signed negative $\log_{10}$ p-values between these two different analyses on the SEA-AD dataset are correlated, we find that many of the statistically significant genes with differential variance are not necessarily the ones with differential mean (Figure 4B).

We also perform a downstream GSEA analysis to understand better the pathways that have an increased variance (i.e., greater loss of coordination) in AD donors. Our GSEA analysis of the SEA-AD cohort illustrates gene sets involved in immune-related functions (Fig. 4C). This finding strongly aligns with the central role of microglia, the brain's resident immune cells, in Alzheimer's Disease (AD) pathology. The significant variability in immune-related gene expression patterns suggests that microglia responses may be highly heterogeneous across AD patients, potentially reflecting different disease states or progression rates. This heterogeneity in immune response could help explain the varying effectiveness of AD treatments across patient populations and supports the growing recognition that AD is not a uniform disease but rather a complex syndrome with multiple contributing factors.

Overall, our results demonstrate a substantial decrease in gene coordination in the microglia of AD donors compared to control donors. These findings complement our previous analyses for differential mean using DESeq2, NEBULA, and eSVD-DE. This demonstrates that important biological insights extend beyond differences in mean expression, and we ask if there could be a general framework that could scan for many types of differences between case and control donors simultaneously. We pursue this using the framework of *differential distributions* in the next section.

Wasserstein-2 decomposition for differential distribution

The Wasserstein-2 distance is a mathematical tool that measures distributional differences between gene expression profiles. This approach offers a more holistic approach to analyzing genomic differences between biological samples. Unlike standard differential expression (DE) methods that focus primarily on mean differences, the Wasserstein-2 distance considers the entire distribution of gene expression levels. This comprehensive perspective enables the detection of subtle yet biologically significant changes that might be overlooked when only mean values are compared. By capturing variations in the full distribution, including aspects like spread and shape, the Wasserstein-2 distance enhances understanding of biological processes and regulatory mechanisms influencing gene expression.

Building upon the strengths of the Wasserstein-2 distance, we have developed a statistical framework to decompose it into three distinct contributions: differential mean, variance, and shape. We use an exact decomposition of Wasserstein-2 distance to investigate whether changes in mean, variance, or shape of the gene expression distributions drive the differences among donors. This decomposition enables us to pinpoint specific contributions:

- **Mean Differences**: Suggest repression or activation of pathways (typically the focus of most studies),
- **Variance Differences**: Suggest changes in gene coordination,
- **Shape Differences**: Suggest alterations in cell state composition.

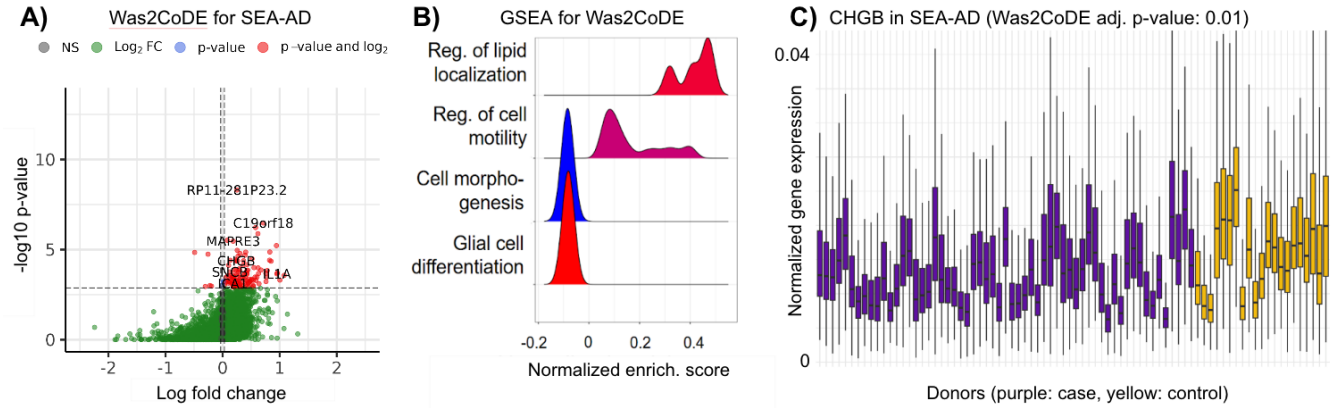

Figure 5. **A**) Volcano plot of the Was2CoDE results. The statistically significant genes with large or small LFCs are marked in red or blue, respectively. Depending on their LFC magnitude, the remaining genes are marked in green or gray. **B**) Normalized enrichment scores (NES) for various cellular processes, including lipid localization, cell motility, cell morphogenesis, and glial cell differentiation. **C**) Box plot of normalized CHGB expression levels across different donors where the Was2CoDE adjusted p-value is 0.01. All plots are based on the SEA-AD dataset.

This decomposition is pivotal as it allows us to dissect the specific factors contributing to the observed differences in gene expression distributions. The differential mean component reflects changes in the average expression levels of genes, indicating potential upregulation or downregulation associated with certain biological conditions or genes. The differential variance component captures changes in gene expression variability, which may suggest alterations in gene regulation, expression noise, or cellular heterogeneity. The differential shape component encompasses more complex distributional changes, such as bimodality, potentially revealing additional insights into cell population dynamics or rare cell states. The appeal of using the Wasserstein-2 distance for cohort-level scRNA-seq analyses can then be summarized as the following: We can develop *one* test that scans for *which* genes have any changes between case and control donors. Then, we can use a mathematical decomposition to help us interpret *how* these genes are specifically different between case and control donors. By unifying the previous differential mean and variance analyses under the Wasserstein-2 framework, we can be assured we leave no biological difference unexplored. We call our method *Was2CoDE* (Wasserstein-2 based Cohort Differential Expression).

Our results of Was2CoDE reveal biological findings recapitulated by other studies but not necessarily found using standard DE methods on scRNA-seq data. For instance, the volcano plot highlights key genes statistically different in distribution between case and control donors (Fig. 5A). This volcano plot has a non-standard skew due to how the LFC is computed. The LFC for Was2CoDE represents the difference between 1) the average Wasserstein-2 distance between a case and a control donor and 2) the average Wasserstein-2 distance between any two cases or any two control donors. The skew originates from the observation that the differences between a case and control donor are often more pronounced than between two donors of the same class. When we perform a GSEA based on the LFC values, we find statistically significant pathways based on Was2CoDE that were not detected by any standard DE method in Figure 3 (Fig. 5B). Furthermore, we can qualitatively inspect what aspect of the distribution drives the statistical significance. For instance, focusing on the CHGB (Chromogranin B), a gene implicated in AD[21], which is detected in our Was2CoDE analysis, we see that the control donors often have a larger expression and a smaller variance when compared to the case donors (Fig. 5C). Our decomposition of Wasserstein-2 allows us to formalize this statement, as we will show next.

While Was2CoDE could reveal unique biological insights, we now set out to ensure that by using the decomposition of the Wasserstein-2 distance, we can recover results that significantly correlate with methods that test for only differential mean, such as DESEq2 with pseudobulking. It helps us understand whether the significant findings from WasCoDE align with those from standard DE methods, which predominantly focus on mean differences. To assess the extent of alignment between WasCoDE's mean contribution and standard DE methods, we posit that if significant genes can indeed be attributed to differences in mean expression, the GSEA results derived from WasCoDE's mean component should correlate with those obtained from existing DE methods. A strong correlation in GSEA results would indicate that the mean differences captured by WasCoDE are consistent with those identified by standard methods, reinforcing the validity of Was2CoDE's decomposition approach.

In Prater's and SEA-AD datasets, the modest positive correlations between WasCoDE and DESeq2 (0.167 and 0.237, respectively) between both method's normalized enrichment scores from GSEA suggest that both methods capture overlapping

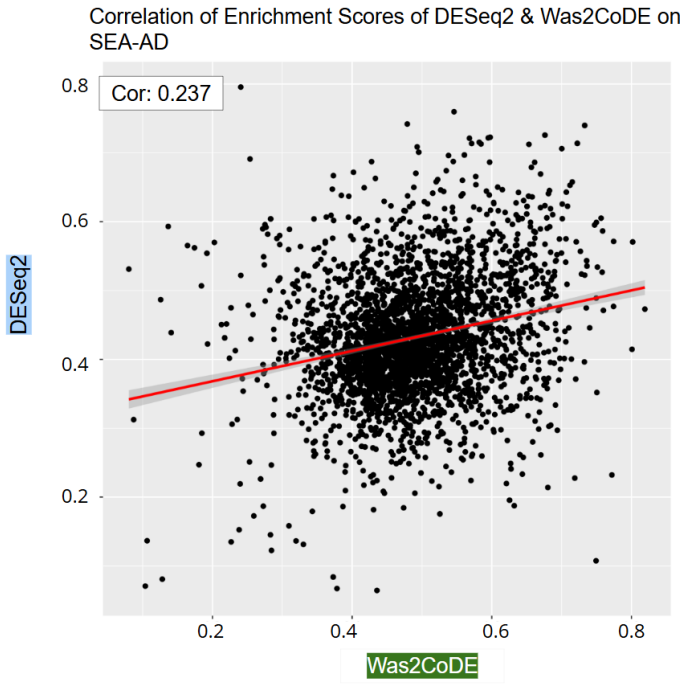

Figure 6. Correlation between DESeq2 (y-axis) and Was2CoDE (x-axis) enrichment scores in the SEA-AD dataset from two GSEAs, where each point represents a biological process pathway.

biological signals (Fig. 6). The SEA-AD dataset, in particular, demonstrates a relatively stronger correlation, reflecting a closer alignment in the biological processes identified by both methods. This higher correlation likely stems from the more controlled structure of the SEA-AD cohort, where donor variability is less pronounced, and mean-level changes dominate as the primary drivers of expression differences. In contrast, the ROSMAP dataset shows a weak and slightly negative correlation (-0.069), which could be driven by the lower number of nuclei per donor. Overall, these results collectively emphasize WasCoDE's complementary value to standard DE methods, validating mean-driven signals while revealing additional layers of biological complexity.

Discussion

In this study, we perform a comprehensive evaluation of differential expression (DE) methods applied to single-nucleus RNA sequencing (snRNA-seq) data from microglia in Alzheimer's disease (AD). By analyzing three distinct datasets (Prater, SEA-AD, and ROSMAP), we uncover insights into the similarities and differences among DE methods. Our findings reveal that while methods such as NEBULA and DESeq2 exhibit high correlations in their results across datasets, certain discrepancies persist. For instance, eSVD-DE can find meaningful pathways across the three datasets that neither NEBULA nor DESeq2 can detect. Looking beyond mean differences, we also find biological insights when analyzing differential variance instead of differential mean expressions. Analyzing the differential variance offers a different perspective on the microglia processes in AD donors. We find many genes with a larger variance in AD donors compared to controls, suggesting that the microglia in AD are undergoing accelerated aging or a loss of coordination. This inspires us to develop Was2CoDE, a method based on the Wasserstein-2 distance, allowing us to identify all distributional differences. We demonstrate that this aids in uncovering significant biological variability that underscores the complexity of AD pathology.

Looking ahead, two key areas merit further exploration. First, a deeper understanding of why DE methods diverge is essential. Specifically, future work should investigate the extent to which factors like varying nuclei counts per donor, microglia subtypes, or intrinsic differences in statistical frameworks contribute to these discrepancies. Such investigations would enhance the reliability and interpretability of DE analyses in single-cell and single-nucleus datasets. Second, the Was2CoDE framework offers a unique opportunity to delve into the underlying drivers of distributional differences. By leveraging its decomposition, future studies could systematically examine how differences in mean, variance, and shape collectively or independently contribute to observed p-values, potentially uncovering new therapeutic targets or biological pathways in AD.

Together, these directions promise to advance our understanding of microglia transcriptomics and their role in AD, paving the way for more nuanced and actionable insights.

Methods

Differential gene expression analysis for differential mean

Our differential gene expression analysis leverages several advanced methods that handle single-nuclei RNA sequencing data in distinct ways. Each method is designed to adjust for confounders while testing for differential expression at the single-nuclei level. All our primary analyses, such as those shown in Figures 2 and 3, adjust for donor's age at death, sex, APOE-ε4 status (i.e., an indicator whether the donor has genetic variant coding for the APOE-ε4 isoform of the APOE gene), and ethnicity, as well as the technical confounders of sequencing batch and post-mortem interval (PMI).

We briefly discuss all the various cohort-level differential expression methods we use in our analysis:

- DESeq2[5] is a widely used method in bulk RNA sequencing adapted for single-nuclei data[6]. It fits a negative-binomial generalized linear model to estimate differences in mean expression between groups while adjusting for size factors and overdispersion in the data. In this paper, we "pseudobulk" the scRNA-seq data before applying DESeq2. This means we combine all the nuclei from each donor, where all the gene transcripts are added together. This process emulates performing bulk RNA-sequencing on a "pure" set of nuclei with the same cell type for each donor.
- NEBULA[3] is a negative binomial mixed model (NBMM) designed specifically for differential expression analysis in cohort-level scRNA-seq datasets. Its key innovation lies in efficiently accounting for the hierarchical structure of single-nuclei data, which typically requires a computationally expensive two-layer optimization in traditional NBMM models due to the intractable marginal likelihoods. NEBULA overcomes this by deriving a closed-form approximation of the marginal likelihood. This analytical approximation eliminates the need for two-layer optimization, enabling significant speed improvements. This innovation allows NEBULA to efficiently account for subject-level and nuclei-level overdispersion, achieving significant computational speedups while maintaining accuracy.
- eSVD-DE[4] tests for differential expression in cohort-level scRNA-seq datasets using a matrix factorization to pool information across genes. This framework enables eSVD-DE to reduce noise and more accurately adjust for confounding covariate effects. In particular, it first estimates a matrix factorization jointly across all the genes to model the scRNA-seq data. Then, it estimates the posterior distribution of each nuclei's gene expression by assessing how well each gene conforms to the statistical model. Finally, for one gene at a time, the method aggregates across all the nuclei originating from each donor, and a hypothesis test is performed by comparing the distribution of gene expression profiles between all the case donors and all the control donors.

Gene-set enrichment analysis (GSEA)

To identify the functional implications of the differential gene expression, we perform Gene-Set Enrichment Analysis (GSEA)[22] using the `clusterProfiler` package from Bioconductor[23]. This approach allows us to assess whether specific gene sets, such as those associated with biological processes or pathways, were significantly enriched among the differentially expressed genes. Throughout the paper, we perform all our GSEA using the `clusterProfiler::gseGO` function, where we quantify the enrichment of all the biological pathways (i.e., "BP") in the `org.Hs.eg.db` Bioconductor package (version 3.18.0, December 2023). In particular, for Figure 3, the GSEA is performed using the LFC of each method, averaged across the three datasets for each DE method.

Details for denoising and normalizing the scRNA-seq dataset

We denoise (i.e. remove the effect of measured confounders from the gene expressions) and normalize (i.e., remove the effect of the sequencing depth) from the scRNA-seq dataset using scVI[24]. The measured confounders include the donor covariates that we have used in our analysis of differential means: donor's age at death, sex, APOE-ε4 status (i.e., an indicator whether the donor has genetic variant coding for the APOE-ε4 isoform of the APOE gene), and ethnicity and the technical confounders of sequencing batch and post-mortem interval (PMI).

Details for the testing the variance

Variance calculations were performed at the gene level for each donor by iterating over genes and computing variances across nuclei. This is performed on the denoised and normalized scRNA-seq dataset via scVI. This process generated a variance matrix, with donors as rows and genes as columns, capturing gene-specific variance profiles across conditions. A Wilcoxon test is applied to the variance matrix, followed by p-value adjustment using the Benjamini-Hochberg method to control for multiple comparisons to evaluate differences in variance between cases and controls. Additionally, the LFC in variance between the two

conditions is calculated to quantify the magnitude of variance shifts for each gene, with results stored with the p-values. For further biological interpretation, GSEA uses the LFC values, focusing on GO Biological Processes, to uncover the functional implications of observed variance differences. These results are shown in Figure 4.

Details for the Was2CoDE

Our differential test method, inspired by IDEAS[7], goes beyond mean differential expression by utilizing the Wasserstein distance to test for differential distribution. This approach allows us to capture not only changes in mean expression across conditions but also significant within-individual variability and broader distributional shifts. This is particularly important for detecting heterogeneity in gene expression patterns that may not be reflected by mean differences alone. We describe each step of our procedure below in each section.

Wasserstein-2 Distance Decomposition

The Wasserstein-2 (i.e., Was2) metric is our foundation for measuring distributional differences between gene expression profiles. For any two donors, we compute the Was2 distance by decomposing it into several key components: location (differences in mean expression), size (differences in variance of expression), and shape (differences in distribution pattern). Specifically, for any two univariate distributions P and Q, the Wasserstein-2 (Was2) distance can be decomposed into location and shape components:

$$W_2^2(P,Q) = (\mu_P - \mu_Q)^2 + (\sigma_P - \sigma_Q)^2 + 2\sigma_P\sigma_Q\rho, \tag{1}$$

where μ_P, μ_Q are the means, σ_P, σ_Q are the standard deviations, and ρ is the correlation coefficient between the optimally coupled random variables under P and Q. The first term, $(\mu_P - \mu_Q)^2$, quantifies differences in mean (i.e., location), while the second term, $(\sigma_P - \sigma_Q)^2$, captures differences in variance (i.e., size). The shape difference metric, computed using quantile correlation ρ, measures the dissimilarity between the distributions' shapes while accounting for potential nonlinear relationships. In our work, P and Q are the distribution of (denoised and normalized) expressions of a particular gene $g \in \{1,\dots,p\}$ between two different donors. This decomposition is also used in waddR[25], another method that tests for differential distributions for scRNA-seq data, but waddR does not focus on cohort-level scRNA-seq analyses.

Computation of donor-pair statistics

Let $x_{g,c}$ denote the normalized and denoised gene expression for gene $g \in \{1,\dots,G\}$ (among a total of G genes) and nuclei (i.e. "cell") $c \in \{1,\dots,m\}$ (among a total of m nuclei). This is performed on the denoised and normalized scRNA-seq dataset via scVI. We compute pairwise metrics between donors through the following procedure. Let $P_{g,i}$ denote the empirical distribution among $\{x_{g,c_1},\dots,x_{g,c_{n_i}}\}$, the gene expression among the n_i nuclei from donor $i \in \{1,\dots,n\}$ (among a total of n donors) for gene g. Likewise, let $P_{g,j}$ be the same as for the donor j. For each donor pair (i,j), we compute several key statistics:

$$\begin{aligned}
\text{Total Wasserstein-2 distance between the two donors}: &\quad W_2^2(P_{g,i},P_{g,j}),\\
\text{Difference in donor means}: &\quad \mu_{g,i} - \mu_{g,j},\\
\text{Difference in donor standard deviations}: &\quad \sigma_{g,i} - \sigma_{g,j}.
\end{aligned}$$

where $\mu_{g,i}$ and $\sigma_{g,i}$ denote the empirical mean and standard deviation in $\{x_{g,c_1},\dots,x_{g,c_{n_i}}\}$. Based on the decomposition (1), we can compute the difference in shape between the two distributions $P_{g,i}$ and $P_{g,j}$. This is equivalent to explicitly computing the ρ term in (1), which can be estimated using correlation between the quantiles of the empirical distributions of $P_{g,i}$ and $P_{g,j}$. We employ kernel density estimation when necessary to smooth the empirical distributions.

Statistical testing

Next, we describe how to use the decomposition (1) to perform a hypothesis test. We categorize donor pairs into three distinct groups. Let $\mathscr{C} \subset \{1,\dots,n\}$ denote the set of case donors and $\mathscr{N} \subset \{1,\dots,n\}$ denote the set of control donors. We define

$$\begin{aligned}
D_{DD} &= \{W_2^2(P_{g,i},P_{g,j}):\ i,j \in \mathscr{C}, i<j\},\\
D_{NN} &= \{W_2^2(P_{g,i},P_{g,j}):\ i,j \in \mathscr{N}, i<j\},\\
D_{DN} &= \{W_2^2(P_{g,i},P_{g,j}):\ i \in \mathscr{C}, j \in \mathscr{N}\},
\end{aligned}$$

where D_{DD} represents within-case Wasserstein distances (where the subscript "D" stands for "dementia"), D_{NN} represents within-control Wasserstein distances (where the subscript "N" stands for "non-dementia"), and D_{DN} represents between-group Wasserstein distances. To ensure balanced comparisons, for each donor i, we randomly sample one case donor $c \in \mathscr{C} \setminus \{i\}$ and one control donor $n \in \mathscr{N} \setminus \{i\}$ for comparison.

For each gene g, we perform a Wilcoxon rank-sum test to assess the significance of distributional differences. Let $D_{\text{within}} = D_{DD} \cup D_{NN}$ denote the pooled within-group distances. The null hypothesis is

H_0 : The distribution of D_{DN} equals the distribution of D_{within}.

The test statistic W is computed as the sum of ranks of D_{DN} elements in the combined ordered sample of D_{within}. We employ a one-sided test to detect significant distributional differences, as we are particularly interested in cases where between-group distances are systematically larger than within-group distances. A large test statistic indicates a meaningful biological variation between conditions. The resulting p-values are adjusted for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). These p-values are shown in Figure 5.

Downstream analysis and biological interpretation

We design a procedure to systematically investigate the relationship between the GSEA results of WasCoDE's mean contribution and those of standard DE analyses. This procedure involves conducting GSEA separately on the significant genes identified by the mean component of WasCoDE and on those identified by existing DE methods. We then perform a comparative analysis of the enrichment scores, p-values, and the specific biological pathways or gene sets highlighted in each analysis. We aim to determine the degree of overlap and divergence between the methods by examining these results.

For each gene g, we compute a mean-contribution score S_g that quantifies the impact of mean differences on the Wasserstein-2 (i.e. Was2) distance,

$$S_g = -\log_{10}(p_g) \times \frac{1}{|\mathscr{C}||\mathscr{N}|} \sum_{\{i\in\mathscr{C}, j\in\mathscr{N}\}} \frac{(\mu_{g,i} - \mu_{g,j})^2}{W_2^2(P_{g,i}, P_{g,j})},$$

where p_g is the Was2CoDE p-value for gene g. This average represents the average ratio of the location component to the total squared Was2 distance across all case-control pairs. Due to the decomposition (1), we are guaranteed that $S_g \in [0,1]$. This score helps identify genes where changes in mean expression primarily drive distributional differences.

We perform two separate GSEAs. One uses the Was2CoDE mean contribution denoted by S_g, and another is the negative $\log_{10}$ p-value of a standard differential expression method, such as from DESeq2 with pseudobulking. We then compute the enrichment of each pathway in each GSEA, regardless of their statistical significance. We compute the correlation of the normalized enrichment score (NES) between the two GSEA results. This correlation between the NES for each GSEA is shown in Figure 6.

References

1. Prater, K. E. *et al.* Human microglia show unique transcriptional changes in Alzheimer's disease. *Nat. Aging* **3**, 894–907 (2023).
2. Das, S., Rai, A. & Rai, S. N. Differential expression analysis of single-cell RNA-seq data: Current statistical approaches and outstanding challenges. *Entropy* **24**, 995 (2022).
3. He, L. *et al.* NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).
4. Lin, K. Z., Qiu, Y. & Roeder, K. eSVD-DE: Cohort-wide differential expression in single-cell RNA-seq data using exponential-family embeddings. *BMC Bioinforma.* **25**, 113 (2024).
5. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
6. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
7. Zhang, M., Liu, S., Miao, Z. *et al.* IDEAS: Individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol.* **23**, 33 (2022).
8. Gabitto, M. *et al.* Integrated multimodal cell atlas of Alzheimer's disease. *Nat. Neurosci.* 1–18 (2024).
9. Sun, N. *et al.* Human microglial state dynamics in Alzheimer's disease progression. *Cell* **186**, 4386–4403 (2023).
10. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).

11. Bennett, D. A. *et al.* Religious orders study and RUSH memory and aging project. *J. Alzheimer's Dis.* **64**, S161–S189 (2018).

12. Ng, A. & Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Adv. Neural Inf. Process. systems* **14** (2001).

13. Abbe, E. Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18**, 1–86 (2018).

14. Wang, F., Mukherjee, S., Richardson, S. & Hill, S. M. High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking. *Stat. Comput.* **30**, 697–719 (2020).

15. Hernando-Herraez, I. *et al.* Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat. Commun.* **10**, 4361 (2019).

16. Levy, O. *et al.* Age-related loss of gene-to-gene transcriptional coordination among single cells. *Nat. Metab.* **2**, 1305–1315 (2020).

17. 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).

18. Leote, A. C., Lopes, F. & Beyer, A. Loss of coordination between basic cellular processes in human aging. *Nat. Aging* **4**, 1432–1445 (2024).

19. Zou, D. *et al.* Single-cell and spatial transcriptomics reveals that PTPRG activates the m6A methyltransferase VIRMA to block mitophagy-mediated neuronal death in Alzheimer's disease. *Pharmacol. Res.* **201**, 107098 (2024).

20. Patel, D. *et al.* Association of rare coding mutations with Alzheimer disease and other dementias among adults of European ancestry. *JAMA Netw. Open* **2**, e191350–e191350 (2019).

21. Marksteiner, J. *et al.* Distribution of chromogranin B-like immunoreactivity in the human hippocampus and its changes in Alzheimer's disease. *Acta Neuropathol.* **100**, 205–212 (2000).

22. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

23. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics: A J. Integr. Biol.* **16**, 284–287 (2012).

24. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

25. Schefzik, R., Flesch, J. & Goncalves, A. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. *Bioinformatics* **37**, 3204–3211 (2021).

Acknowledgements

We thank Suman Jayadev and Eardi Lila for helpful analysis ideas helped shape this work.

Author contributions statement

K.Z.L. and K.E.P. planned the analyses. W.T.Z. coded the Was2CoDE. W.T.Z. and K.Z.L. performed the analysis. W.T.Z. led the writing, but all authors contributed to the writing. All authors reviewed the manuscript.

Code and analysis reproducibility

Code used in this work is publicly available. The method implementation can be found at https://github.com/TatiZhang/Was2CoDE. For reproducing our analysis and results, see https://github.com/linnykos/Was2CoDE_analysis.