Large-scale snRNA-seq meta-analysis of microglia role in Alzheimer's disease across statistical methods

Wenjing Tati Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2025

Reading Committee
Kevin Z. Lin, Chair
Katherine E. Prater

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Large-scale snRNA-seq meta-analysis of microglia role in Alzheimer's disease across statistical methods

Wenjing Tati Zhang

Chair of the Supervisory Committee:
Kevin Z. Lin

Department of Biostatistics

Microglia orchestrate complex neurodegeration processes that drives Alzheimer's disease (AD), yet their transcriptional signatures remain inconsistently reported across single-nucleus RNA-seq studies. We analyze three pre-frontal-cortex cohorts (Prater, SEA-AD, ROSMAP; ranging from 22 to 345 donors) with five differential-expression (DE) pipelines and introduce Was2CoDE – a Wasserstein-2–based test that partitions donor-to-donor differences into mean, variance and shape components. Three principal findings emerge. First, study design matters: the rigorously curated SEA-AD cohort reproducibly recovers the highest fraction of literature-validated microglia pathways, and power scales chiefly with the number of donors, not nuclei or read depth. Second, among DE frameworks, the matrix-factorization approach eSVD-DE delivers the most consistent gene- and pathway-level signals across independent datasets. Third, we shed light on the opportunity to discover underlying microglia mechanisms by analyzing differential distributions, which is broader than differential mean expression. Specifically, Was2CoDE uncovers distributional shifts missed by mean-centric tests, revealing variance-driven dysregulation in immune and cell-motility programs and highlighting genes such as ARHGEF3, CD9, and SASH1 that escape standard DE thresholds. Together, these results provide quantitative guidance for cohort design, benchmark analytic robustness and supply an open-source tool for full-distribution inference. By integrating method, design and distributional insights, our framework advances the search for microglia therapeutic targets in AD.

Large-scale snRNA-seq meta-analysis of microglia role in Alzheimer's disease across statistical methods

Wenjing Tati Zhang[1], **Jinqiu Turbo Du**[1], **Yihan Chen**[1], **Suman Jayadev**[2], **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 orchestrate complex neurodegeration processes that drives Alzheimer's disease (AD), yet their transcriptional signatures remain inconsistently reported across single-nucleus RNA-seq studies. We analyze three pre-frontal-cortex cohorts (Prater, SEA-AD, ROSMAP; ranging from 22 to 345 donors) with five differential-expression (DE) pipelines and introduce Was2CoDE – a Wasserstein-2–based test that partitions donor-to-donor differences into mean, variance and shape components. Three principal findings emerge. First, study design matters: the rigorously curated SEA-AD cohort reproducibly recovers the highest fraction of literature-validated microglia pathways, and power scales chiefly with the number of donors, not nuclei or read depth. Second, among DE frameworks, the matrix-factorization approach eSVD-DE delivers the most consistent gene- and pathway-level signals across independent datasets. Third, we shed light on the opportunity to discover underlying microglia mechanisms by analyzing differential distributions, which is broader than differential mean expression. Specifically, Was2CoDE uncovers distributional shifts missed by mean-centric tests, revealing variance-driven dysregulation in immune and cell-motility programs and highlighting genes such as *ARHGEF3*, *CD9*, and *SASH1* that escape standard DE thresholds. Together, these results provide quantitative guidance for cohort design, benchmark analytic robustness and supply an open-source tool for full-distribution inference. By integrating method, design and distributional insights, our framework advances the search for microglia therapeutic targets in AD.

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. However, microglia's diverse and intricate functions still need to be better understood, complicating efforts to develop targeted therapeutics. Understanding the transcriptomic phenotypes 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 reliably detect microglia-specific transcriptional differences between AD and non-AD donors, since microglia are much smaller in size than neurons and 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, two major limitations of existing single-nuclei workflows remain, 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 regarding whether or not these methods should yield consistent biological insights, particularly as the number of donors in the cohort expands. A critical distinction among these methods lies in how they model how nuclei from the same donor have more similar transcriptomic profiles than nuclei from

different donors. 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] and edgeR[6] can be adapted to single-nuclei data through pseudo-bulking, where nuclei from each donor are aggregated[7]. 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. While studies such as[8] have computationally benchmarked various DE methods, these (1) have not been extensively studied for complex settings among large cohorts of human donors, and (2) did not tease apart what aspects of an experimental design or statistical analyses would yield more result in more generalizable findings.

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[9] 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.

In this study, we systematically evaluate the performance and consistency of multiple differential expression methods across varying cohort sizes using single-nucleus RNA sequencing data from human microglia in AD. Using three well-characterized single-nucleus RNA-seq cohorts of human pre-frontal microglia (Prater, SEA-AD, and ROSMAP), we (1) disentangle how experimental design factors—donor sample size, nuclei depth, and sequencing depth—govern cross-dataset reproducibility of mean-centric DE pipelines; (2) benchmark four representative DE frameworks (DESeq2[5], NEBULA[3], eSVD-DE[4], edgeR[6], and MAST[10]) to identify the method that yields the most generalizable signals; and (3) introduce Was2CoDE, a Wasserstein-2–based decomposition that partitions donor-to-donor transcriptional differences into orthogonal mean, variance, and shape components, thereby exposing biologically meaningful distributional shifts missed by conventional tests. By coupling systematic down-sampling experiments with cross-cohort validation against an external silver-standard gene set, we provide quantitative guidance on where to invest experimental resources and demonstrate that integrating mean-based and distribution-based analyses delivers a more comprehensive picture of microglia dysregulation in 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[11], and the ROSMAP cohort[12]. 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 is demonstrated in Figure 1, where we see a substantially higher number of microglia per donor than in other datasets in this paper. We refer to this dataset as the "Prater dataset."
- The SEA-AD (Seattle Alzheimer's Disease) dataset[11] consists of 80 donors (59 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[13], 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,

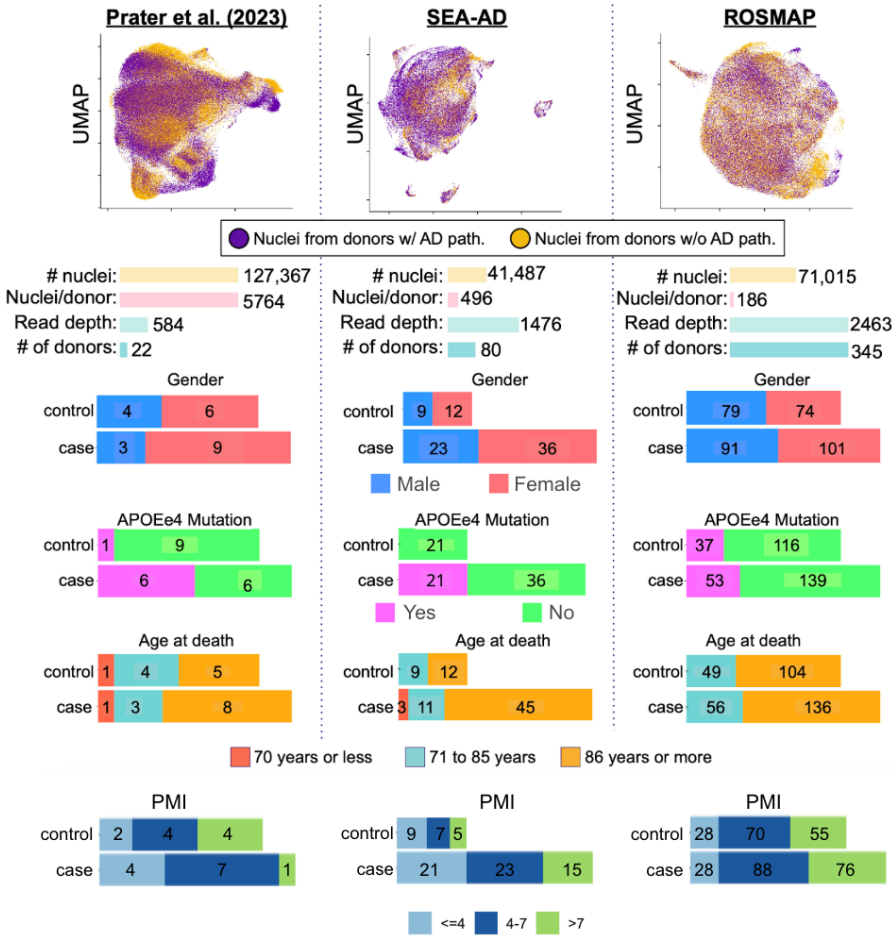

Figure 1. Overview of the three snRNA-seq datasets to study Alzheimer's disease (AD) research: Prater et al. (2023)[1], SEA-AD[11], and ROSMAP[12]. UMAP visualizations (top) display the transcriptional profiles from the three datasets, with nuclei colored by donor AD pathology status (purple: donors 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, and the range of post-mortem interval (PMI). 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).

this study has a high number of microglia per donor, given the modestly large number of donors.

- The ROSMAP dataset[12] consists of 345 donors from the ROSMAP study[14] (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.

Investigating the relationship between the 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*[15–17]. 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,

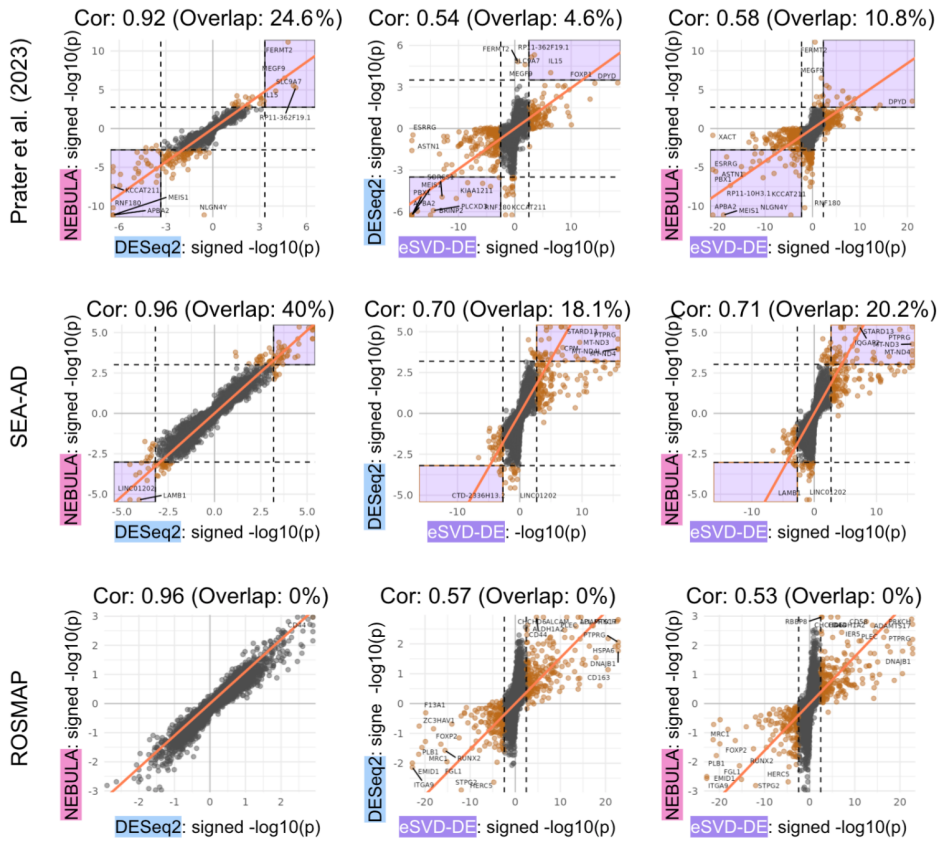

Figure 2. **Comparison among datasets and statistical methods**: Results are shown for three datasets (Prater, SEA-AD, and ROSMAP), analyzed using three statistical methods (NEBULA, DESeq2, and eSVD-DE). Each scatter plot displays signed -$\log_{10}$ (p-values) between pairs of methods, with correlation coefficients at the top of each panel. Each point is a gene. The directionality of sign of the -$\log_{10}$ p-value is based on sign of the log-fold change (i.e., the expression of case donors minus control donors). 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. The solid orange line denotes leading the principal component in the respective plot. The dotted lines denote the multiple testing threshold on a log scale, and the purple regions denote genes that cross the multiple-testing threshold in both methods. 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").

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. While our primary analyses focus on the three methods, we also perform edgeR and MAST and selectively showcase it when it aids in the interpretation. 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, SEA-AD, and ROSMAP. 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. NEBULA shows consistently high correlation with DESeq2 (Pearson correlation = 0.92-0.96), while correlations between other method pairs are more moderate (Pearson correlation = 0.53-0.71), suggesting strong agreement between NEBULA and DESeq2 methodologies across different datasets. Finally, we note that the correlation between the methods' estimated LFC is often higher than the correlation between the methods' signed negative $\log_{10}$ p-values.

We also noted that the original analysis of the microglia in the ROSMAP by Sun et al. (2023)[12] discovered 1,542 DE

genes via MAST, but our analysis using DESeq2 and NEBULA yielded no DE genes. To understand this discrepancy between our analysis and the original analysis, we applied MAST on the ROSMAP data in two ways, where we adjust for the same covariates as we have discussed. We observed that when we applied MAST disregarding the donor hierarchical structure (i.e., comparing all the nuclei from any case donor to all the nuclei from any control donor), we find 1,457 DE genes (FDR $<$ 0.05). This set highly overlaps with the DEs initially reported in Sun et al. (2023). However, when we applied MAST where we account for which nuclei originates from which donors via random effects in the intercepts, there were no DE genes that passed the multiple-testing threshold. This fact demonstrates the necessity to account for donor structure in future differential expression testing for snRNA-seq data. By properly accounting for the donor structure, we surmise that the number of donors is what drives the statistical power, not the number of nuclei. However, since our analysis on ROSMAP seemed to have less statistical power than on SEA-AD, we dive deeper into this premise in remainder of the paper to understand the nuances of this premise.

Overall, 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. Furthermore, DESeq2 and NEBULA are not able to find any significant genes in the ROSMAP cohort. In the following sections to come, we investigate the drivers of these findings—what considerations about the experimental design or statistical analysis enable more generalizable findings?

Investigating experimental design considerations that yield more generalizable findings

A challenge in snRNA-seq studies is that lists of differentially expressed (DE) genes often diverge from one dataset to the other. Much of this variation stems from diverse design choices — brain-region sampling, donor composition, nuclei-per-donor, library chemistry, or sequencing depth—that distinguish among different snRNA-seq experiments that investigate the role of microglia in AD. These design choices can influence which genes appear DE, sometimes more than the downstream statistical method itself. Our goal in this section is therefore to identify which experimental designs yield DE results that most faithfully reproduce external, independently validated microglia signatures regardless of which method those DE results come from. A design that recovers known biology across methods is likely to support findings that generalize to future cohorts.

We assessed the extent to which the DE results from each of the four methods – DESeq2, eSVD-DE, NEBULA, and edgeR – agree with the gene-level findings from a silver-standard gene set. Specifically, to construct this silver-standard gene set to be maximally stringent, we retained only the genes in both lists —the "Mathys & Zhou" (MZ) intersection, comprising n = 72 genes which are enriched for 123 GO terms, as an external benchmark. These two papers[18,19] were used in the HuMicA[20] analysis that constructs a microglia atlas across many neurodegenerative diseases, more than specifically AD. We will call the genes and pathways derived from & Zhou as the MZ genes and MZ pathways, respectively, which focuses on AD. Because part of the ROSMAP cohort was also profiled in the Mathys study, any gene or pathway that appears only in Mathys's could be replicated in ROSMAP simply by shared individuals. By construction, even though ROSMAP and Mathys share participants, only those genes that also reached significance in Zhou – a completely separate cohort – enter our evaluation. Restricting to the intersection ensures that every retained gene has been replicated across at least two discovery cohorts, thereby neutralizing this source of confounding.

We first asked whether the MZ pathways are reproducibly recovered across three AD datasets—Prater, SEA-AD, and ROSMAP—regardless of the DE algorithm applied (DESeq2, NEBULA, eSVD-DE, or edgeR). For each dataset–method combination, we investigated the subset of MZ genes identified as DE (FDR $<$ 0.05) and recorded whether the corresponding MZ pathways were deemed significant. Across the 123 MZ pathways, SEA-AD reproduces 37 pathways (30.1%), whereas ROSMAP and Prater recover only 11 (8.9%) and 18 (14.6%) pathways, respectively. Notably, none of the MZ pathways were detected across all three datasets, while 44 pathways (35.8%) were found in at least one dataset, and 79 pathways (64.2%) were not detected in any dataset. Within each dataset, we examined method-specific performance and found that SEA-AD consistently identified more MZ pathways regardless of the method used (32 to 38 pathways). In contrast, method performance varied more in the Prater and ROSMAP datasets, with eSVD-DE showing the highest detection in both (8 and 12 pathways, respectively). To provide a method-agnostic view of each dataset's performance, we also tested for enrichment of the MZ gene list using Fisher's exact test on the DE gene sets derived from each method and summarized the resulting p-values. Further, we compared gene set enrichment results from DESeq2 and eSVD within the SEA-AD and ROSMAP datasets (Figure 3A,B). In SEA-AD, DESeq2 and eSVD showed strong agreement (Pearson correlation = 0.96), with many of the 80 MZ pathways (out of 3097 total tested) jointly detected as significant and concentrated in the top right quadrant, indicating concordant results. In contrast, the ROSMAP dataset demonstrated weaker agreement (Pearson correlation = 0.75), with only 79 MZ pathways evaluated and fewer showing consistent significance across methods. Many MZ pathways in ROSMAP were marginally significant or uniquely detected by one method, suggesting a greater degree of method-specific variability in this dataset.

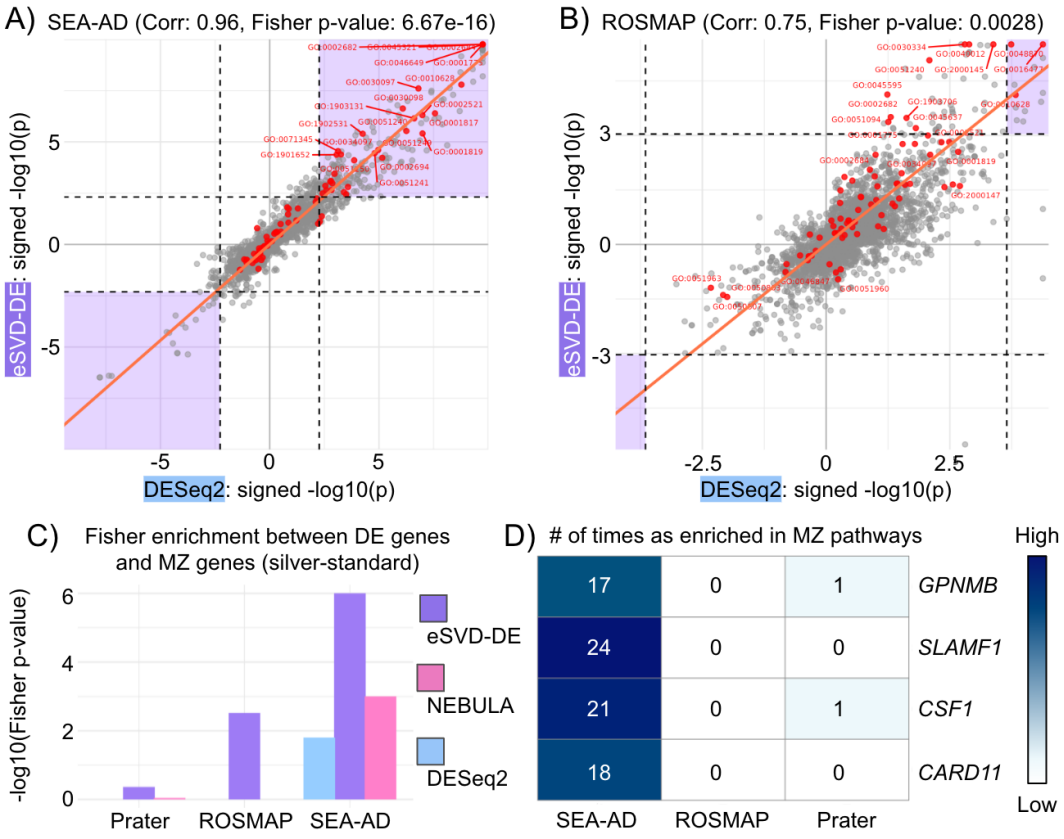

Figure 3. **Comparison of different methods on one dataset**: **(A)** and **(B)**: scatter plots comparing signed -$\log_{10}$ p-values from DESeq2 (x-axis) and eSVD (y-axis) for pathway-level GSEA results in the SEA-AD and ROSMAP datasets, respectively. Each dot represents a GO pathway, with values signed according to the direction of enrichment. Red dots represent the MZ pathways. The MZ pathways are labeled in red. The remaining plotting aesthetic follow Figure 2. **(C)**: Fisher's exact test results quantifying enrichment of significant pathways in the MZ pathways, stratified by method and dataset. Bars show -$\log_{10}$ p-value for each test. **(D)**: Heatmap showing occurrence of four representative MZ genes (rows) across datasets (columns) within pathways identified as significant by any method. The color intensity corresponds to the number of pathways containing the gene.

To quantify the amount of overlap across the three cohorts, we used Fisher's exact test at both the gene and pathway levels: first comparing each method's FDR-significant genes with the 72 Mathys–Zhou (MZ) Alzheimer's-microglia genes, and then assessing enrichment of each method's significant pathways within the MZ pathway set (Figure 3C). At the gene level, SEA-AD displayed the strongest concordance: eSVD-DE achieved a 6.0-fold enrichment ($p = 1.0 \times 10^{-6}$), while NEBULA and DESeq2 reached 3.0-fold ($p = 1.0 \times 10^{-3}$) and 1.8-fold ($p = 1.6 \times 10^{-2}$) enrichment, respectively. ROSMAP showed a weaker signal: eSVD-DE reached 2.5-fold enrichment ($p = 3.0 \times 10^{-3}$), whereas NEBULA returned no FDR-significant overlap and DESeq2 was non-significant ($p = 1$). Prater offered little concordance, with eSVD-DE showing a log10 enrichment of only 0.36 ($p = 0.43$) and both NEBULA ($p = 0.91$) and DESeq2 ($p = 1$) essentially null. The same hierarchy emerged at the pathway level: eSVD-DE produced the strongest –$\log_{10}$ p-values, reaching 5.9 in SEA-AD, 2.3 in ROSMAP, and 0.6 in Prater, whereas DESeq2 and NEBULA showed only modest enrichment in SEA-AD and ROSMAP and negligible signal in Prater. Collectively, these results reveal two clear patterns. First, eSVD-DE consistently outperforms NEBULA and DESeq2 in recovering AD-relevant microglia signatures. Second, recovery is highly dataset-dependent, with SEA-AD providing the most conducive experimental context, ROSMAP intermediate, and Prater the least—underscoring the pivotal role of study design in rediscovering the MZ pathways.

Our analysis of the top MZ pathways revealed three distinct functional groups across datasets. The first group, comprising immune regulation pathways such as regulation of leukocyte proliferation (GO:0070663) and B cell activation (GO:0042113), was detected strongly in both SEA-AD and Prater (4 detections each) but absent in ROSMAP. The second group, including cell signaling pathways such as cellular response to cytokine stimulus (GO:0071345) and positive regulation of cytokine production (GO:0001819), was consistently detected in SEA-AD (4 detections each) with variable presence in the other datasets.

The third group contained cell migration and motility pathways, including cell migration (GO:0016477), regulation of cell motility (GO:2000145), and regulation of locomotion (GO:0040012), which showed balanced detection between SEA-AD and ROSMAP.

Of the 72 Mathys-Zhou (MZ) genes that recur across datasets, four—*CX3CR1*, *HSPA1A*, *SLC11A1*, and *SPP1*—are consistently differentially expressed in Prater, SEA-AD, and ROSMAP, each long associated with microglia activation or stress responses in neuro-degeneration. These genes populate many of the top GO terms we recover: in SEA-AD they appear in 26 of the 38 reproduced MZ pathways, spanning immune-activation (e.g. chemokine-mediated signalling, GO:0070098), stress-response (protein refolding, GO:0042026), and phagocytic modules (phagosome organisation, GO:1904874); their coverage drops to roughly half that number in ROSMAP's 12 significant pathways and still further in Prater's 11. To probe the microglia programme more deeply, we next tallied the genes most frequently encountered within a curated subset of immune-related pathways (Figure 3D), identifying *GPNMB*, *SLAMF1*, *CSF1*, and *CARD11* as the four dominant drivers. In SEA-AD these genes are richly represented—*SLAMF1* appears in 24 significant pathways, *CSF1* in 21, *GPNMB* in 17, and *CARD11* in a comparable number—whereas their presence in ROSMAP is sparse and in Prater largely absent. Together, the two four-gene panels reinforce the same message: SEA-AD offers the most faithful recovery of canonical microglia biology across analytic methods, while ROSMAP and especially Prater provide progressively weaker and more method-dependent signals.

Our comparative analysis demonstrates that experimental design impacts the reproducibility and generalizability of findings in microglia genomics. Among the datasets we've analyzed, the SEA-AD dataset consistently outperformed other experimental designs in detecting validated microglia pathways across all analytical methods. This enhanced performance was not method-dependent but rather a direct consequence of robust experimental design choices. Future studies investigating microglia biology in neuro-degenerative diseases should adopt design principles similar to those employed in SEA-AD. Building cohorts along these lines will make downstream analytic choices less pivotal and yield DE findings that translate more reliably across laboratories, technologies, and disease contexts.

Investigating statistical methods that yield more generalizable findings

Another interest is to evaluate which statistical methods produce findings that are most consistent across independent datasets. Employing methods that yield consistent results across diverse datasets reduces the likelihood that conclusions are driven by cohort-specific artifacts, since therapeutic hypotheses build on scientific findings. Our assessment was predicated on two fundamental criteria: (1) methodological robustness should manifest as correlated statistical outcomes between independent datasets, and (2) effective methods should consistently identify established biological signals regardless of dataset origin. We examined the correlation of significance values ($-\log_{10}$ p-values) between datasets for each method. Beyond technical reproducibility, we evaluated biological relevance by comparing the MZ genes and pathways identified as previously described. Using Fisher's exact test, we quantified the enrichment of the overlapping significant genes within literature-validated gene sets.

eSVD-DE demonstrated substantially higher correlation coefficients across the three datasets compared to DESeq2 and NEBULA, suggesting greater consistency in detecting differential expression signals more broadly. Our pathway analysis revealed several key Alzheimer's disease mechanisms identified by eSVD-DE in both datasets that were not detected by either DESeq2 or NEBULA. We compared pathway-level enrichment results derived from the eSVD-DE method between the Prater and SEA-AD cohorts (Figure 4A) and observed a moderate correlation in statistical significance across datasets, with a substantial number of pathways found to be significant in both (Pearson correlation = 0.53), indicating that many biological signals are preserved despite differences in cohort characteristics. The gene-level statistics also exhibited moderate cross-dataset correlation (Pearson correlation = 0.33). Furthermore, pathways found to be significant in both datasets were strongly enriched for those reported in prior literature, with a Fisher's exact test yielding a p-value of 5.7×10^{-5}. These findings suggest that the eSVD-DE approach captures signals that generalize across datasets despite differences in cohort composition and technical processing.

Fourteen immune-related GO pathways are captured by eSVD-DE in at least two dataset-pairs, whereas no other method surpasses one hit for most of them. The list includes key proliferation terms—regulation of leukocyte proliferation (GO:0070663), B-cell activation (GO:0042113), regulation of mononuclear- and lymphocyte-cell proliferation (GO:0032944, GO:0050670)—as well as broader programmes such as positive regulation of gene expression (GO:0010628) and regulation of immune-system process (GO:0002682). Motility-oriented pathways, such as cell migration and its regulatory variants (GO:0016477, GO:0030334, GO:0040012, GO:2000145), and cytokine-centric terms (GO:0001817, GO:0001819) crossed the multiple-testing threshold. For every one of these 14 pathways, eSVD-DE detects each one in two different datasets, indicating reproducible detection in at least two of the three cohort comparisons. By contrast, DESeq2 and edgeR reach a count of two in only a handful of cases, and NEBULA drops to one (or zero) for most pathways. The pattern is clearest in the two key terms – regulation of leukocyte proliferation and B-cell activation – where eSVD-DE, DESeq2, edgeR, and NEBULA each score two

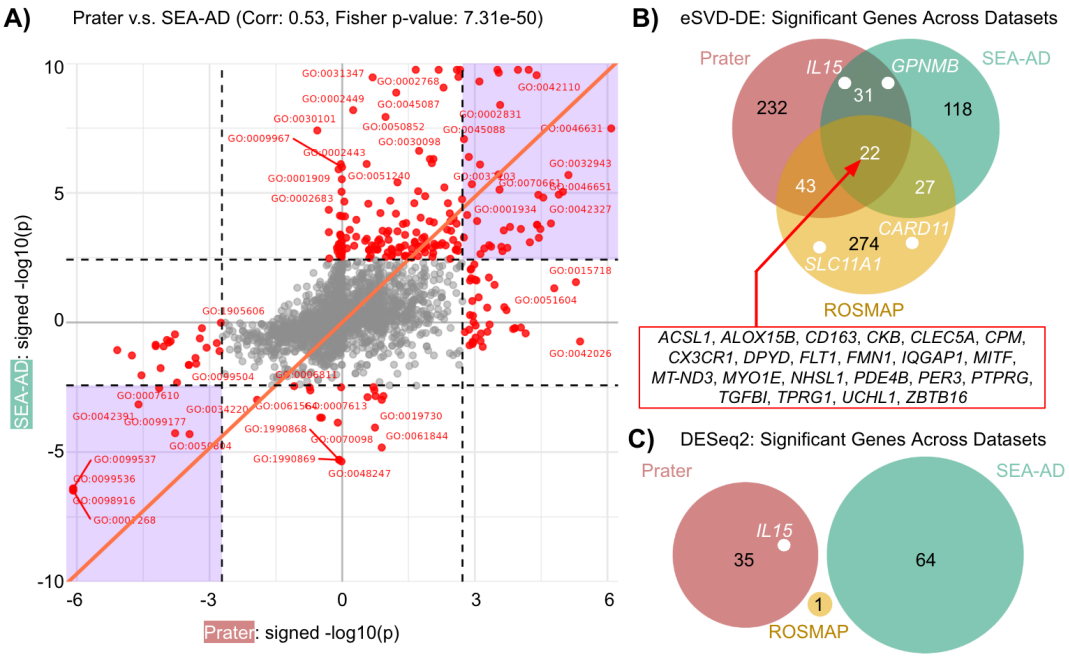

Figure 4. **Cross-dataset reproducibility of pathway- and gene-level findings**: **(A)**: Scatter plot comparing signed -$\log_{10}$ p-values of gene set enrichment results from the eSVD-DE method across the Prater and SEA-AD datasets. Each point represents a GO biological pathway; red points indicate pathways significant (FDR < 0.05) in at least one of the two datasets. The remaining plotting aesthetic follow Figure 2. **(B)**: Venn diagram of genes found significant by eSVD-DE across the Prater, SEA-AD, and ROSMAP datasets. The intersection includes 22 genes consistently identified in all three datasets; *IL15*, *GPNMB*, *CARD11*, and *SLC11A1* are labeled in white. **(C)**: Venn diagram showing DESeq2-derived significant genes across the same datasets.

hits, yet only eSVD-DE maintains that performance across all remaining pathways.

We narrowed the top 20 MZ pathways that are consistently found significant across datasets to 11 immune-related pathways that govern leukocyte proliferation, cytokine production, and cell motility, then interrogated how reliably DESeq2, eSVD-DE, and NEBULA reproduced these gene–pathway signals in Prater, SEA-AD, and ROSMAP. For every pairwise dataset comparison we extracted the genes shared by a pathway in both datasets and tallied how many times each gene recurred in the three possible overlaps (i.e., between Prater and SEA-AD, Prater and ROSMAP, or SEA-AD and ROSMAP). *IL15* emerges as the most reproducible, appearing in 8 of 9 data set pair overlaps - six from eSVD-DE and one each from DESeq2 and NEBULA - and participating in six of 11 tracked immune pathways, including regulation of leukocyte proliferation (GO:0070663) and regulation of mononuclear-cell proliferation (GO:0032944). Its strongest support appears in Prater (eSVD-DE $p = 4.2 \times 10^{-7}$, DESeq2 $p = 9.3 \times 10^{-5}$, NEBULA $p = 1.4 \times 10^{-5}$). *GPNMB* is recovered in six overlaps (four via eSVD-DE) and maps to four of the same immune pathways (eSVD-DE $p = 1.2 \times 10^{-6}$ and 1.3×10^{-4}) but not in ROSMAP. Of the MZ genes, only *CARD11* and *SLC11A1* are detected, both uniquely captured by eSVD-DE; *CARD11* reached $p = 1.2 \times 10^{-5}$ in SEA-AD, while *SLC11A1* peaked in ROSMAP (2.2×10^{-7}), potentially reflecting dataset-specific microglia activation uniquely captured by eSVD-DE.

At the gene level, we examined the overlap of significant findings across Prater, SEA-AD, and ROSMAP. Using eSVD-DE, we identified 22 genes that were consistently significant across all three datasets (Figure 4B). We also highlighted the genes we found (*IL15*, *GPNMB*, *CARD11*, and *SLC11A1*) in this set which are implicated in neuroinflammatory processes and Alzheimer's pathology. In contrast, DESeq2 did not produce overlapping gene-level findings, with the vast majority of significant genes being unique to a single dataset (Figure 4C). This limited reproducibility underscores the method's sensitivity to cohort-specific noise and its reduced capacity to identify consistent biological signals across heterogeneous single-nucleus RNA-seq datasets.

Based on these multiple lines of evidence, our findings suggest that eSVD-DE may be particularly well-suited for capturing subtle but biologically meaningful transcriptional changes associated with complex neurodegenerative processes when compared to comparable statistical methods. These include a higher cross-cohort agreement in signed negative $\log_{10}$ p-values, higher recovery of immune-motility pathways in at least two datasets, and consistency of rediscovering key genes.

This enhanced generalizability is pivotal in the context of single-nuclei RNA-sequencing of human brain tissue, where biological and technical variability can often obscure true disease-associated signals.

Downsampling experiment to isolate the impacts of experimental design and statistical method

Based on our findings in the previous sections, we now investigate which experimental considerations contributed the most to the power of the statistical method. This gets to the heart of our investigation – when future researchers design new snRNA-seq analyses of microglia to study AD, what is the "most important" aspect to consider? Is it more important to prioritize recruiting more post-mortem donors into the study design, obtaining more nuceli per donor, or sequencing each nuclei more deeply? This is critical question to investigate due to the costly nature of snRNA-seq experiments.

To appreciate the different experimental design aspects that contribute to the power of statistical analysis, we performed an *in silico* downsampling experiment across all three datasets. For each dataset, we varied the number of nuclei per donor, the sequencing depth per donor, and the number of donors across 10 different levels across 10 trials. Specifically, when downsampling the number of nuclei per donor, we uniformly at random sampled nuclei per donor if that donor contributed more nuclei than the desired amount. When downsampling the sequencing depth per nuclei, we proportionally reduced the read count for each nuclei across all the genes, injecting minimal randomness as needed to break ties when needed. When downsampling the number of donors, we uniformly at random sampled donors based on sex and case-control status to preserve the balance between AD and non-AD donors jointly with the balance between the donor sexes. We hold the other aspects fixed whenever downsampling one aspect.

Overall, we found that the number of donors was biggest contributor to statistical power, followed by the number of nuclei per donor, and that the loss of power was more gradually reduced using eSVD-DE when compared to DESeq2 and NEBULA. Our findings are shown in Figure 5, which shows the percentage of DE genes found in each *in silico* downsampled dataset, relative to the original number of DE genes found in that dataset-method combination. Focusing on eSVD-DE, we notice that in the SEA-AD dataset, the number of DE genes dropped by roughly half when there are only 22 donors (as opposed to the original 80 donors), and in the ROSMAP dataset, likewise, by roughly a quarer when are only 22 donors (opposed to the original 345 donors). Notably, the standard error bars is considerably larger when downsampling donors compared to downsampling the other two aspects. Overall, our findings demonstrate that regardless of which statistical method is used, the number of donors drive the majority of the statistical power. Relating back to Figure 2, we suspect the reason that SEA-AD recovered more DE genes compared to ROSMAP despite having not as many donors is primarily driven by the donor recruitment. We expand upon this point in more detail in the Discussion.

Investigating changes beyond changes in mean expression

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[21–25]. Regarding AD, we hypothesize that donors with AD pathology have microglia that "age faster" or undergo accelerated disorganization[26]. This can statistically manifest as donors with AD pathology have a higher variance in certain genes than donors without AD pathology.

Building on IDEAS[9], an established method for testing differential distributions between donor groups, we extend the method by incorporating a Wasserstein-2 distance decomposition that partitions inter-donor gene-expression differences into mean, variance, and distributional-shape components, thereby providing a more nuanced analysis than traditional mean-centric approaches. 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. 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:

- **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.

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

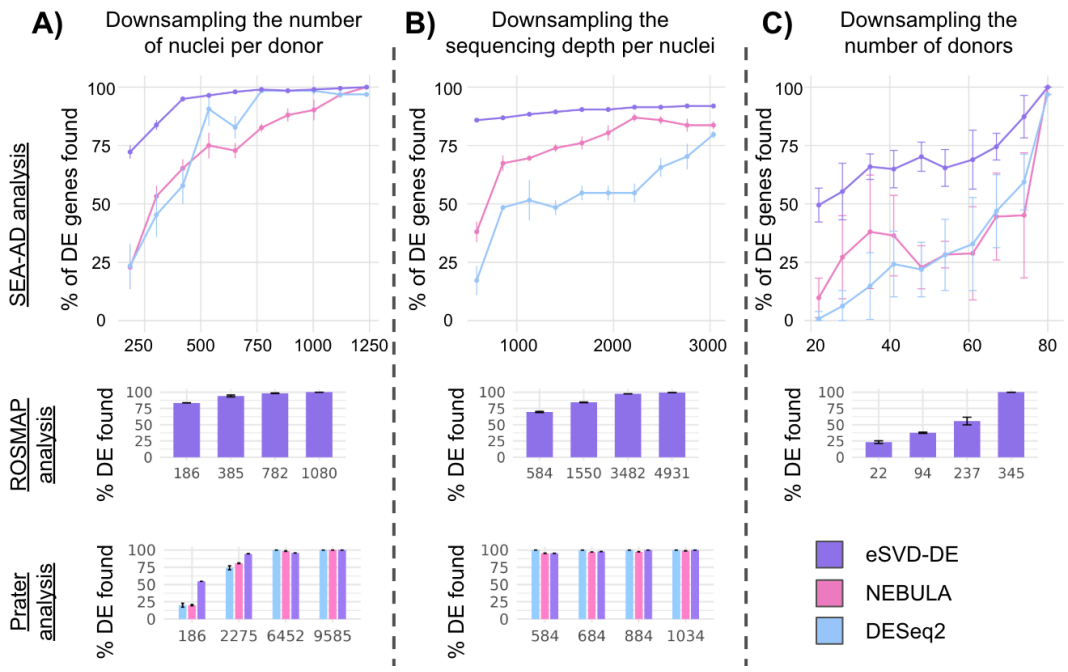

Figure 5. **Downsampling of the datasets, based on number of nuclei per donor, sequencing depth, or number of donors**: **(A)**: The number of nuclei per donor is downsampled from a range starting from the maximum of number of nuclei across any donor to the small median sequencing depth across the three datasets (186). The y-axis represents the percentage of DE genes found relative to the DE genes found in the original analysis of that dataset (i.e., no downsampling), where the value denoted is the median percentage across 10 trials, with error bars based on by $\pm$IQR/2 (inter-quantile range). We only applied eSVD-DE to the ROSMAP dataset since there were no DE genes discovered in the original analysis for DESeq2 or NEBULA. **(B)**: Similar to (A), except the sequencing per donor is downsampled from a range starting from 90% quantile among all nuclei to the small median sequencing depth across the three datasets (584). **(C)**: Similar to (A), except the number of donors is downsampled from a range starting from the total number of donors to the smallest number of donors across all three datasets (22). The color assignment of each method is the same as in Figure 2.

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. We call our method *Was2CoDE* (Wasserstein-2 based Cohort Differential Expression), and its computation is overviewed in Figure 6A. The output of Was2CoDE for each gene is, as we will describe, a log-fold change for Wasserstein distance or any of the three components or a p-value summarizing the total differential distribution between case and control donors.

We applied our Was2CoDE framework to determine if there were genes with identify differential distributional patterns that are not captured by previous differential expression (DE) methods. We apply this preliminary analysis on the SEA-AD dataset, since we have demonstrated in Figure 3 that this dataset exhibits stronger method concordance and higher generalizability. Our broad analysis revealed a set of microglia-relevant genes such as *ARHGEF3*, *CD9*, and *SASH1* that exhibit large differential shape and variance components under Was2CoDE, but were missed entirely by all three DE methods (DESeq2, NEBULA, and eSVD-DE) (Figure 6B). These genes participate in biological processes implicated in immune signaling, actin cytoskeletal regulation, and cell adhesion, and are enriched in microglia activation and phagocytic pathways. To evaluate statistical significance for these select genes, we applied a post-hoc PERMANOVA test using $500,000$ permutations to compute a p-value for how the distribution for this gene differs between case and control donors. The resulting p-values were 2.6×10^{-5} for *ARHGEF3*, 1.4×10^{-5} for *CD9*, and 3.3×10^{-4} for *SASH1*, indicating robust distributional shifts. In contrast, the smallest p-values obtained from any of the three DE methods for these genes were 0.11 (*ARHGEF3*, Wilcoxon test), 0.18 (*CD9*), and 0.15 (*SASH1*) – all above the nominal 0.05 threshold. These results demonstrate that traditional DE frameworks lack power to detect these subtle but biologically meaningful patterns, calling for the need for distributional approaches in single-nuclei analysis. Notably, Was2CoDE pinpoints why: its size term captures heightened donor-to-donor dispersion, while the shape term isolates heavier tails and latent multimodality in case donors—distributional nuances that elude standard DE analyses.

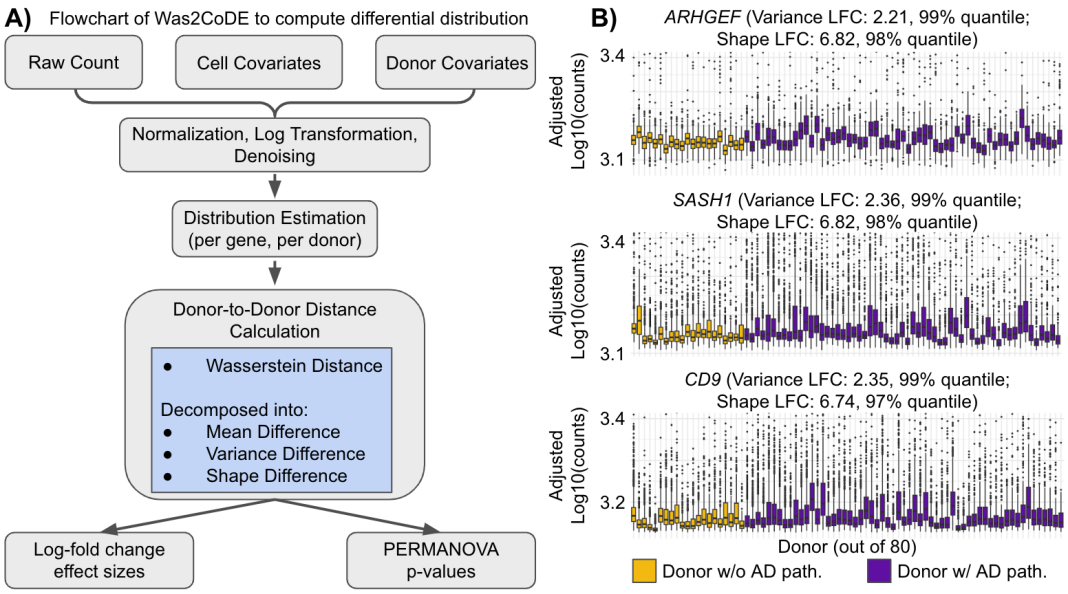

Figure 6. Was2CoDE method to understand differential distribution: **(A)** Workflow of Was2CoDE, illustrating the major steps of the Was2CoDE pipeline. The analysis begins with raw gene expression counts along with nuclei- and donor-level covariates. After normalization, log transformation, and denoising, expression distributions are estimated for each gene within each donor. Donor-to-donor distances are then computed using the Wasserstein-2 distance, which captures full distributional differences. This distance is further decomposed into interpretable components: mean difference, variance difference, and shape difference. These decomposed components are used to derive two main statistical outputs: (1) log-fold change effect sizes between donor groups, and (2) PERMANOVA p-values for significance testing. **(B)**: Donor-level expression patterns of example genes with similar means but distinct distributional shifts in the SEA-AD dataset. We report log-fold changes for the variance and shape components for these genes, along with the quantile, which denotes how many genes have a smaller log-fold change.

Using the full Was2CoDE decomposition, we compute the Was2CoDE log-fold change transcriptome-wide and re-tested every gene pathway for based significant enrichment for the variance and shape components. We compared Was2CoDE's shape and size GSEA results against conventional DE pipelines across both SEA-AD and Prater datasets. As shown in Figure 7, Was2CoDE identified 465 and 312 significant pathways in SEA-AD and Prater, respectively—far exceeding the 442 and 165 pathways captured by the traditional DESeq2-eSVD-NEBULA ensemble. Among the 25 pathways uniquely shared between Was2CoDE across both datasets but not found using any of the DE methods (Figure 7, highlighted), many relate to transcriptional regulation, RNA processing, and cell proliferation – biological themes frequently implicated in microglia responses to neurodegeneration. For example, several RNA metabolic processes including negative regulation of RNA biosynthetic process, mRNA metabolic process, and regulation of transcription by RNA polymerase II suggest nuanced shifts in transcriptional control not captured by mean-based DE methods. Other recurrent pathways – regulation of cellular response to stress, pattern recognition receptor signaling, and small GTPase-mediated signal transduction – are consistent with microglia activation and innate immune signaling. This could suggest that Was2CoDE reveals several biologically significant pathways that were undetectable through differential expression methods tested on the mean.

These findings support the view that Was2CoDE detects subtle but functionally relevant perturbations, particularly in stress-adaptive and proliferative programs, that remain elusive to conventional frameworks focused on mean expression alone. Because these findings recur in two independent datasets, they may reflect consistent but relatively subtle alterations that mean-centric tests overlook. Our results suggest that incorporating variance and shape information can complement traditional differential-expression analyses without overstating the evidence.

Discussion

Our systematic meta-analysis delivers four principal insights. First, among the three cohorts, the SEA-AD design reproducibly recovers the largest fraction of literature-validated microglia signatures, irrespective of the downstream algorithm. Second, across methods, the matrix-factorization framework eSVD-DE yields the most consistent gene- and pathway-level signals between independent datasets. Third, power analyses based on extensive in-silico down-sampling show that the number of donors—not nuclei per donor or sequencing depth—is the dominant driver of statistical yield. Lastly, by de-

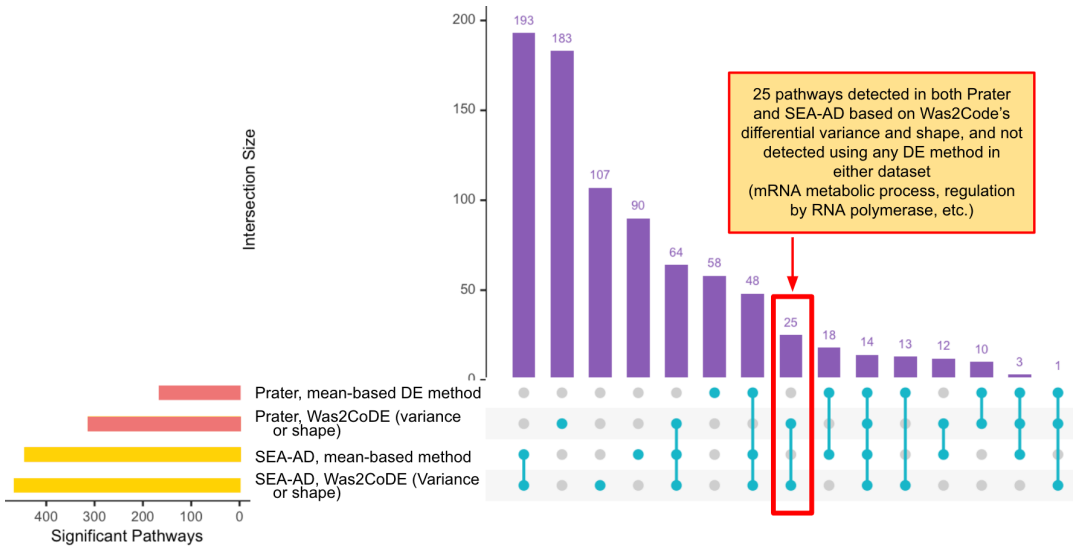

Figure 7. Overlap of significant pathways identified by Was2CoDE and other differential expression methods across datasets: UpSet plot summarizing the intersections of significant GO biological process pathways detected by four analysis pipelines: NEBULA, eSVD, and DESeq2 in both the SEA-AD and Prater datasets, and Was2CoDE (variance & shape components) in both datasets. Horizontal bars on the left indicate the total number of significant pathways identified by each method-dataset combination. Vertical bars show the size of each pathway intersection across selected combinations, with filled blue dots below each bar denoting which methods contributed to that intersection. The 25 pathways uniquely identified by Was2CoDE in both datasets (highlighted bar) include some of the following pathways: mRNA metabolic process, regulation of transcription by RNA polymerase II, regulation of cellular response to stress, cell surface pattern recognition receptor signaling pathway, positive regulation of locomotion, cellular response to endogenous stimulus, regulation of endothelial cell proliferation, negative regulation of cell motility, negative regulation of cell migration, negative regulation of epithelial cell proliferation, cell morphogenesis, negative regulation of signaling, and small GTPase-mediated signal transduction.

composing Wasserstein-2 distances, Was2CoDE uncovers distributional shifts (variance and shape) that escape mean-centric differential-expression tests, highlighting non-trivial microglial dysregulation in Alzheimer's disease.

SEA-AD's improved generalizability likely stems from one of two design choices. First, the consortium used a rigorous tissue quality control, such as ensuring a small post-mortem interval, high RNA integrity and uniform pH. These can contribute to increasing the signal-to-noise at the donor level. Second, the consortium used a purposeful sampling across the full neuropathological continuum—balanced controls, intermediate cases and advanced ADNC. This ensured dense coverage of disease progression and boosts power for intermediate phenotypes. Collectively, these features produce cleaner, more comparable transcriptional readouts across analytic pipelines and external datasets. This can explain our overall findings on why our statistical analyses on the SEA-AD dataset had more power when compared to the ROSMAP dataset, despite the ROSMAP having more donors.

We found that eSVD-DE outperform other DE approaches, and we hypothesize this stems from eSVD-DE jointly factoriings the entire expression matrix while respecting the nested nuclei-within-donor structure. By pooling low-rank signal across thousands of genes, the method amplifies weak but coherent donor-level effects, attenuates sparsity-driven noise and preserves degrees of freedom for covariate adjustment. This combination yields higher cross-dataset correlations in signed $-\log_{10} p$ values and recovers a larger share of immune-motility pathways implicated in microglia activation. The framework thus provides a blueprint for next-generation DE tools that integrate global structure with hierarchical sampling.

The study's strength lies in its breadth and consistency: we re-analyse three large, brain-region-matched snRNA-seq cohorts with five complementary DE frameworks, apply uniform covariate models and benchmark all outputs against an external silver standard. The same workflow can be extended to other cell types, modalities (snATAC, spatial transcriptomics) and neurodegenerative diseases, offering a scalable template for evaluating design and method choices in single-nuclei cohort studies.

Several limitations warrant caution. First, the datasets profiled partially overlapping but not identical microglia sub-states, so residual biological heterogeneity may confound cross-cohort comparisons, although within-dataset method evaluations remain unaffected. Second, Prater and SEA-AD recruit donors from the same geographical area and partially overlapping

donors, which may inflate apparent concordance; furthermore, the researchers collecting and analyzing these two datasets are naturally expected to follow similar protocols, explaining why there was higher alignment between Prater and SEA-AD when compared to ROSMAP. Third, SEA-AD's balanced pathology sampling could itself elevate reproducibility, leaving open whether its advantage derives from design rigour or case–control ratio. Notably, the 80 donors in the SEA-AD cohort has a higher proportion of cases than controls (see Figure 1). Future work with additional, equally controlled cohorts will clarify these issues and further refine best-practice guidelines for cohort-scale single-nuclei studies.

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, 3, and 4, 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[7]. 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 nuclei 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.
- edgeR[6] is a bulk RNA sequencing where we also pseudobulk the scRNA-seq data in our work here. edgeR models the count via a generalized linear model, uses empirical Bayes techniques to moderate dispersion estimates across genes, and uses a likelihood ratio test to test for differential expression.
- MAST[10] is a two-part hurdle model, consisting of a logistic regression to model the probability of gene detection (i.e., whether a gene is expressed) and a Gaussian linear model to characterize the distribution of non-zero expression values. Additionally, MAST incorporates random effects to account of which nuclei originates from which donors.

Each dataset was analyzed with the 5000 highly variable genes.

Gene-set enrichment analysis (GSEA)

To identify the functional implications of the differential gene expression, we perform Gene-Set Enrichment Analysis (GSEA)[27] using the `clusterProfiler` package from Bioconductor[28]. 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).

Details for the Was2CoDE

Our differential test method, inspired by IDEAS[9], 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. It closely follows the workflow performed in IDEAS[9]. Let $x_{g,c} \in \{0,1,2,\dots,\}$ denote raw sequencing counts 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).

Step 1: Gene Expression Normalization

To account for technical covariates such as sequencing depth, the function normalizes expression values on a per-gene, per-donor basis. For each gene and individual, Was2CoDE identifies nuclei belonging to that individual and extracts the raw gene expression data. It then performs a log transformation and fits a linear regression model using specified nuclei-level covariates such as sequencing depth or other factors, regressing log-transformed expression values (i.e., $\log_{10}(x_{g,c}+0.5)$) against the log-transformed per-nuclei normalization variable (e.g., $\log_{10}(\sum_{g=1}^{G} x_{g,c})$ for each nuclei c). Residuals from this model capture the variation in gene expression not explained by technical factors. A fixed intercept correction, based on the median covariate value across all nuclei, is added back to stabilize the scale across donors. The result is a list of adjusted expression distributions for each gene and donor, which we denote as $\tilde{x}_{g,c}$ for nuclei c and gene g.

Next, donors are grouped based on the phenotype variable into two groups (e.g., AD cases and controls). The function then computes pairwise distances between these residualized donor distributions for each gene. For each gene, it constructs a three-dimensional array to store pairwise comparisons between all individuals (number of donors, by number of donors, by the four components of Was2CoDE). These comparisons capture the four distinct components along the third dimension of this array: the overall Wasserstein distance, the location difference (reflecting changes in mean expression), the size difference (capturing variance changes), and the shape difference (representing alterations in distribution pattern).

Step 2: Wasserstein distance and its 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 distance can be decomposed into location and shape components:

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

where μ_P, μ_Q are the means, σ_P, σ_Q are the standard deviations, and ρ_{PQ} 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_{PQ}, 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[29], another method that tests for differential distributions for scRNA-seq data, but waddR does not focus on cohort-level scRNA-seq analyses.

Was2CoDE incorporates procedures for handling missing data during distance computations in the presence of sparsity, a common feature in single-nuclei RNA-sequencing datasets. When calculating pairwise Wasserstein distances between donors, missing expression values for a given gene are excluded from the empirical distributions used in the estimation. To prevent instability due to small sample sizes, donor pairs with insufficient observed values (below a user-defined threshold) are omitted from the analysis for that gene.

Step 3: Computation of donor-pair statistics

We compute pairwise metrics between donors through the following procedure based on the adjusted expressions $\tilde{x}_{g,c}$. Let $P_{g,i}$ denote the empirical distribution among $\{\tilde{x}_{g,c_1},\dots,\tilde{x}_{g,c_{n_i}}\}$, 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:

$$\begin{aligned}
\text{Wasserstein-2 distance between the two donors :} \quad & D^{(1)}_{g,(i,j)} = W_2^2(P_{g,i}, P_{g,j}), \\
\text{Difference in donor means :} \quad & D^{(2)}_{g,(i,j)} = \mu_{g,i} - \mu_{g,j}, \\
\text{Difference in donor standard deviations :} \quad & D^{(3)}_{g,(i,j)} = \sigma_{g,i} - \sigma_{g,j}, \\
\text{Difference in donor distribution's shape :} \quad & D^{(4)}_{g,(i,j)} = 2\sigma_{g,i}\sigma_{g,j}\rho_{g,i,j},
\end{aligned}$$

where $\mu_{g,i}$ and $\sigma_{g,i}$ denote the empirical mean and standard deviation in $\{\tilde{x}_{g,c_1}, \ldots, \tilde{x}_{g,c_{n_i}}\}$. 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 $\rho_{g,i,j}$ term in (1), which can be estimated using correlation between the quantiles of the empirical distributions of $P_{g,i}$ and $P_{g,j}$.

While the above formula computes the difference in any distributional component between any two donors, we now quantify the extent of distributional shifts between cases and controls in aggregate for as effect size, akin to a log fold change. This is defined for each gene g and any component (i.e., $k \in \{1, \ldots, 4\}$, where $k = 1$ is for Wasserstein distance, $k = 2$ is for difference in mean, etc.). Let A (for "Alzheimer's pathology" i.e., case donors) denote donors with pathology for Alzheimer's disease, and N (for "no pathology", i.e., control donors) denote denotes with no pathology for Alzheimer's disease. Let $\mathscr{A} \subset \{1, \ldots, n\}$ denote the set of case donors and $\mathscr{N} \subset \{1, \ldots, n\}$ denote the set of control donors. Inspired by the test statistic for a Welch's T-test (i.e., two-sample T-test with unequal variances), we define the effect size akin to a log-fold change (LFC) as:

$$\text{LFC}^{(k)}_g = \frac{D^{(k;AN)}_g}{\sqrt{\left(\frac{\sigma^{(k;AA)}_g}{\sqrt{n_A}}\right)^2 + \left(\frac{\sigma^{(k;NN)}_g}{\sqrt{n_N}}\right)^2}} \tag{2}$$

where, $n_A = |\mathscr{A}|$ and $n_N = |\mathscr{N}|$ are the number of case and control donors, $D^{(k;AN)}_g$ is the average difference in $D^{(k)}_{g,(i,j)}$ for an case donor i and control donor j, defined as

$$D^{(k;AN)}_g = \frac{1}{n_A n_N} \sum_{i \in \mathscr{A}, j \in \mathscr{N}} D^{(k)}_{g,(i,j)},$$

and $\sigma^{(k;AA)}_g$ and $\sigma^{(k;NN)}_g$ is the standard deviation of difference among case or control donors, respectively. That is, for example,

$$\sigma^{(k;AA)}_g = \text{standard deviation}\left(\left\{D^{(k)}_{g,(i,j)} \,:\, i, j \in \mathscr{A} \text{ and } i \neq j\right\}\right)$$

and similarly defined for $\sigma^{(k;NN)}_g$.

In short, this effect size standardizes the average pairwise distance between case and control donors using the pooled standard error from within-group variability. For each gene g and Wasserstein component, it outputs a unitless effect size analogous to a z-score, indicating how much more different the cases and controls are relative to the expected within-group noise. Higher values of $\text{LFC}^{(k)}_g$ reflect stronger between-group signal for that distribution component $k \in \{1, \ldots, 4\}$.

Step 4: Statistical testing

Next, we describe how to use the decomposition (1) to perform a hypothesis test. We employ a permutation-based multivariate analysis of variance (PERMANOVA) approach to assess statistical significance. By using a permutation-based approach, we avoid making strong statistical assumptions about how the gene expressions relate to the case-control status of donors. This procedure works in the following way: for each gene g, the original Wasserstein distance matrix is used to calculate an F-statistic comparing within-group distances (e.g., case-case, control-control) to between-group distances (case-control), denoted as F_g to assess statistical significance. By avoiding parametric distributional assumptions, the test remains valid for the Wasserstein-2 distance matrices we compute for each gene. The calculation of F_g is as follows: For every gene g, let $\left(D^{(1)}_{g,(i,j)}\right)_{1 \leq i,j \leq n}$ be the $n \times n$ donor–donor squared distance matrix. The pseudo-F statistic used in the permutation test is

$$F_g \;=\; \frac{\left(\text{SST}_g - \text{SSW}_g\right) \times (n-2)}{\text{SSW}_g}, \tag{3}$$

where, letting $G_k = \mathscr{N}$ for $k = N$ and $G_k = \mathscr{A}$ for $k = A$,

$$\text{SST}_g = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} D^{(1)}_{g,(i,j)}, \quad \text{and} \quad \text{SSW}_g = \sum_{k\in\{A,N\}} \frac{1}{n_k}\sum_{i\in G_k}\sum_{j\in G_k} D^{(1)}_{g,(i,j)}.$$

To generate a null distribution of F_g, we following PERMANOVA procedure by performing b permutations (default: $B = 999$) by randomly shuffling the group labels. This preserves the distance matrix structure but reassigns all the n donors to either case or control For each permutation, a new F-statistic is calculated, denoted as $F_g^{(b)}$ for permutation $b \in \{1,\ldots,B\}$. The empirical p-value is then computed as:

$$p\text{-value} = \frac{\left(\sum_{b=1}^{B} \mathbb{1}[F_g^{(b)} \geq F_g]\right) + 1}{B+1},$$

where $\mathbb{1}[x]$ denotes the indicator function that has value 1 when the event x is true.

For datasets with many donors, such as our analysis of ROSMAP with 345 donors, there is a high computational cost of compute $D^{(1)}_{g,(i,j)}$ for every 2 pairs of donors $(i,j) \in \{1,\ldots,n\}^2$. As such, we propose to use a computational sub-sampling where we only compare each donor i to K other case and K other control donors. Specifically, we randomly sample K case donor $d \in \mathscr{A} \setminus \{i\}$ and K control donor $d \in \mathscr{N} \setminus \{i\}$ for comparison. Then, in the calculation of the F_g or $F_g^{(b)}$, we only utilize the values of $D^{(1)}_{g,(i,j)}$ that were calculating (omitting any value in the summation that we did not compute and then re-appropriately adjusting the normalization terms $1/n$ or $1/n_k$).

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. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
7. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
8. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
9. 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).
10. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 1–13 (2015).
11. Gabitto, M. *et al.* Integrated multimodal cell atlas of Alzheimer's disease. *Nat. Neurosci.* 1–18 (2024).
12. Sun, N. *et al.* Human microglial state dynamics in Alzheimer's disease progression. *Cell* **186**, 4386–4403 (2023).
13. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
14. Bennett, D. A. *et al.* Religious orders study and RUSH memory and aging project. *J. Alzheimer's Dis.* **64**, S161–S189 (2018).

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

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

17. 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).

18. Mathys, H. *et al.* Single-cell transcriptomic analysis of alzheimer's disease. *Nature* **570**, 332–337 (2019).

19. Zhou, Y. *et al.* Human and mouse single-nucleus transcriptomics reveal trem2-dependent and trem2-independent cellular responses in alzheimer's disease. *Nat. medicine* **26**, 131–142 (2020).

20. Martins-Ferreira, R. *et al.* The human microglia atlas (humica) unravels changes in disease-associated microglia subsets across neurodegenerative conditions. *Nat. Commun.* **16**, 739 (2025).

21. 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).

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

23. 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).

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

25. Upadhya, S., Klein, J. A., Nathanson, A., Holton, K. M. & Barrett, L. E. Single-cell analyses reveal increased gene expression variability in human neurodevelopmental conditions. *The Am. J. Hum. Genet.* **112**, 876–891 (2025).

26. Dillman, A. A. *et al.* Transcriptomic profiling of the human brain reveals that altered synaptic gene expression is associated with chronological aging. *Sci. Reports* **7**, 16890 (2017).

27. 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).

28. 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).

29. 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 Eardi Lila for helpful analysis ideas helped shape this work. This work was supported by the University of Washington's Royalty Research Fund (A208363).

Author contributions statement

KZL and KEP planned the analyses. WTZ and KZL coded the Was2CoDE. WTZ, JTD, YC, and KZL performed the analysis. WTZ, SJ, KEP, and KZL interpreted the results. WTZ and KZL 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.