Cohort level differential distributional analysis for studying microglia in Alzheimer's via single-cell RNA-seq

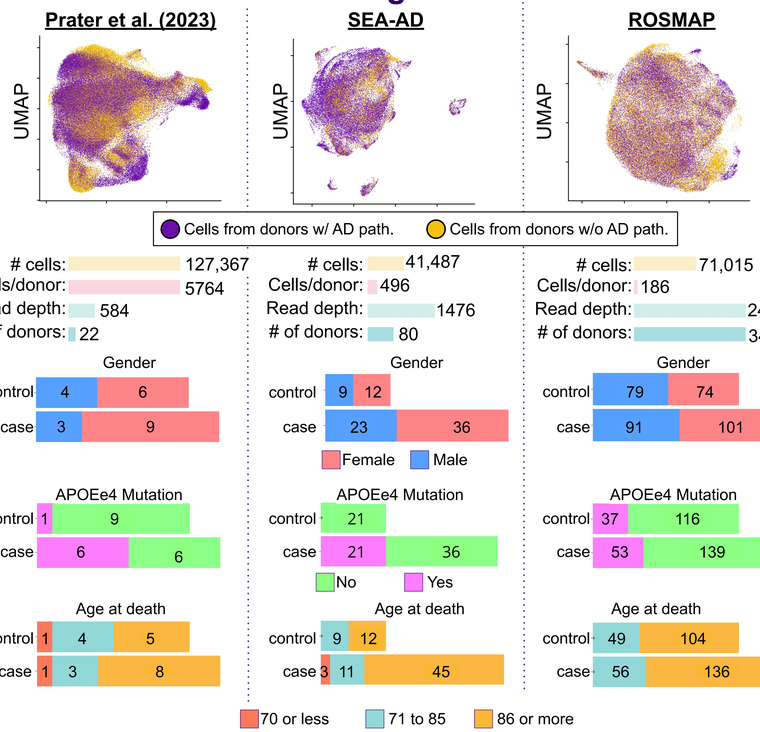
Wenjing (Tati) Zhang[1]; Katherine E. Prater[2]; Kevin Z. Lin[1]

[1]University of Washington (Biostatistics), [2]University of Washington (Neurology)

Introduction

1. AD is an epidemic that affects cognitivity of elderly people.
2. Microglia have been implicated in the progression of Alzheimer's disease. They release inflammatory mediators, aberrantly phagocytose neurons, and facilitate in tau protein spread.
3. Single cell RNA-seq at different donor's brains at varying AD pathology allows us to study the molecular mechanisms of microglia and its impacts on AD.
4. To quantify differences in gene expression for microglia between AD & non-AD donors that is recapitulated across multiple cohorts.

Three human microglia dataset of AD

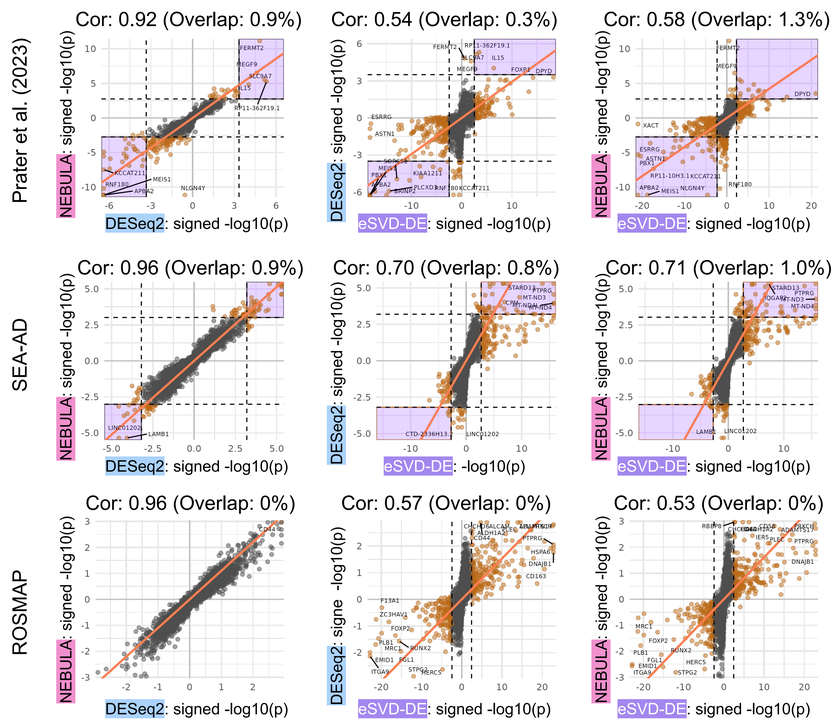


Methods for testing mean difference from cohort data

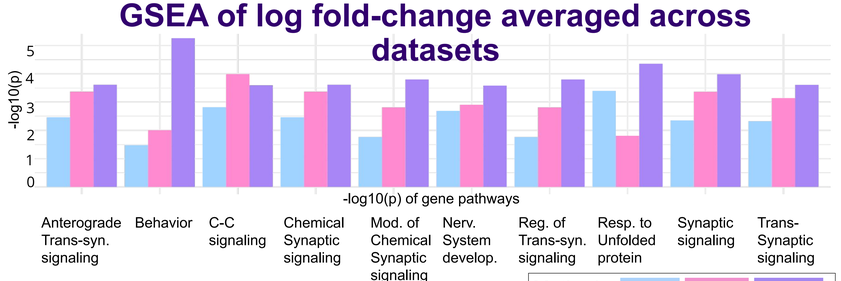
- **DESeq2:** Pseudo-bulk method, summing all cells from each donor prior to adjustment.
- **NEBULA**: Negative binomial mixed model to account for donor covariates, using computational tricks for fast performance and accurate inference.
- **eSVD-DE**: Matrix factorization which pools information across genes, removing confounding covariate effects.

We adjust for Sex, Age at Death, Sequencing Batch, Race, Post Mortem Interval, APOE.4.

Comparisons among different methods



GSEA of log fold-change averaged across datasets



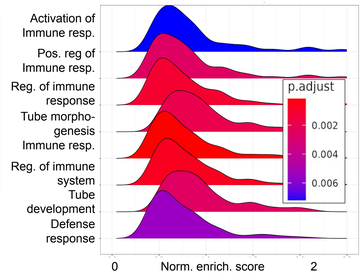
Takeaways:

- NEBULA & DESeq2 are highly correlated among datasets. However, we do not observe any significant genes in ROSMAP.
- eSVD-DE consistently shows neurosignaling pathways with GSEA.
- eSVD-DE shows the most promising enrichment among 3 methods, suggesting that it might be more reliable when analyzing cohort-level scRNA-seq data.

Looking beyond differences in mean

- Gene expression heterogeneity is hypothesized to increase will aging. Is it possible that microglia from AD donors "age faster" than from other donors?
- To explore this possibility, we compute the variances of each gene within each donor and test for differences in variances.

$H_0 : \mathbb{V}(\text{gene}_j \text{ in case}) = \mathbb{V}(\text{gene}_j \text{ in control})$

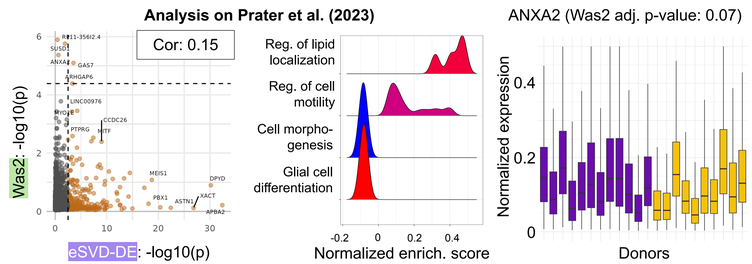


Our method to model all types of distributional differences

Mathematical fact: The Wasserstein distance between two distributions can be decomposed in the % contribution due to shifts in mean, variance, or shape:

$$\text{Wasserstein}^2(X,Y) = \underbrace{(\mu_x - \mu_y)^2}_{\text{Diff. in mean}} + \underbrace{(\sigma_x - \sigma_y)^2}_{\text{Diff. in variance}} + \underbrace{2\sigma_x\sigma_y(1-\rho_{xy})}_{\text{Diff. in shape}}$$

1. **Adjust for confounders:** Regress out the effects of confounding variables on scRNA-seq data (using scVI).
2. **Measure distance:** For each gene, compute the Wasserstein-2 distance between every pair of two donors.
3. **Compute p-values:** Perform a Wilcoxon test between the Wasserstein-2 distance between {one case to one control} vs. {two cases, or two controls}.



Conclusion

- Different statistical methods find similar results in differential gene expression in means between AD & non-AD donors.
- Gene variability tells us there is significant difference in variance between AD & non-ad, serving as evidence that the coordination of gene expression within these cells diverges in Alzheimer's disease. This indicates that the regulatory networks within AD microglia, or other brain cells, may be disrupted or altered.
- We developed a cohesive framework that unifies the diverse findings produced by the current differential expression methods, using the Wasserstein-2 distance.

References: (Datasets): Prater et al. (Nature Aging, 2023), SEA-AD: Gabitto et al. (Nature Neuroscience, 2024), ROSMAP: Sun et al. (Cell, 2023). **(Methods):** DESeq2: Love et al. (Genome Biology 2014), NEBULA: He et al. (Communications Biology 2021), eSVD-DE: Lin et al. (BMC Bioinformatics, 2024), scVI: Lopez et al. (Nature Methods, 2018)