Impact of Cross-Cohort Heterogeneity on Differential Gene Expression and Feature Selection in Alzheimer's Disease

GitHub repository: https://github.com/ylzzzzz/625-Final-Project

Jiamin Zou, Yalei Zhao, Nick Treloar

1 Abstract

Alzheimer's disease is associated with widespread alterations in gene expression, and bulk gene expression studies have been widely used to investigate disease-related molecular changes in human samples [1]. To improve statistical power and generalizability, data from multiple independent cohorts are often integrated; however, such integration is challenged by cross-cohort heterogeneity arising from differences in brain region or tissue source, experimental platforms, and preprocessing pipelines [2]. In this study, we investigate the impact of cross-cohort heterogeneity on differential gene expression analysis and exploratory feature selection in Alzheimer's disease. Using normalized bulk gene expression data from four independent cohorts, we first performed differential expression analyses within each cohort to characterize cross-cohort heterogeneity. We conducted pooled differential expression analyses on the merged dataset both without and with cohort adjustment, and complemented these analyses with UMAP visualizations to assess global expression structure and batch-driven clustering. Finally, robustness of gene–disease associations was assessed by examining the stability of exploratory feature selection in the merged dataset before and after cohort adjustment. Cohort adjustment substantially altered differential expression results and reduced batch-driven clustering patterns. These findings emphasize the need to account for cross-cohort heterogeneity and to assess robustness when integrating gene expression data across cohorts or brain regions in Alzheimer's disease studies.

2 Introduction

Alzheimer's disease is a progressive neurodegenerative disorder marked by cognitive decline and extensive transcriptional changes, and its molecular characterization has relied heavily on large-scale gene expression studies. However, reproducibility and cross-study comparability remain challenging, as reported disease-associated genes often vary substantially between studies due to differences in brain region or tissue source, experimental platforms, preprocessing pipelines, etc [2]. Cross-cohort heterogeneity introduces biological variation and technical batch effects that can distort gene–disease associations [3]. Integrating multiple AD cohorts enables assessment of the sensitivity of disease-associated signals to cohort structure and batch effects. Given that disease status is known for all samples and cohort sample sizes are limited, the primary goal of this study is statistical inference: identifying genes whose expression differs between Alzheimer's disease and control samples and assessing the sensitivity of these associations to cohort structure and batch effects.

In this study, we examined gene–disease association analyses in Alzheimer's disease using four cohorts. We first characterized differential gene expression between AD patients and controls

within individual cohorts. We then evaluated the impact of cross-cohort integration by performing pooled analyses with and without cohort adjustment. Finally, we assessed the robustness of disease-associated signals by examining consistency of feature selection.

3 Methods

Data Sources and Preprocessing: Normalized bulk gene expression data were obtained from the GenoTEX benchmark dataset provided through Kaggle (https://www.kaggle.com/datasets/haoyangliu14/genotex-llm-agent-benchmark-for-genomic-analysis/data#dataset), which compiles publicly available gene expression studies from the Gene Expression Omnibus (GEO). All expression values were log-transformed and normalized by the original studies, and no additional normalization was performed. We initially identified eight Alzheimer's disease–related cohorts, each provided as a separate CSV file corresponding to a single GEO series (e.g., GSEXXXXXX.csv). Within each cohort file, rows represent individual samples indexed by sample identifiers, and columns include sample-level metadata and gene expression measurements. Metadata variables consisted of a binary Alzheimer's disease indicator (1 = Alzheimer's disease, 0 = control), age, and gender, while the remaining columns corresponded to normalized gene expression values for a common set of genes across cohorts. Data were read using the `data.table` package in R, which was benchmarked to provide faster input performance than alternative data-loading methods.

The cohorts differed in experimental platform, study design, and brain region of origin, including middle temporal gyrus, frontal cortex, and temporal cortex. To ensure adequate statistical power for differential expression analysis, cohorts with fewer than 50 total samples were excluded from pooled analyses. As a result, three cohorts (GSE137202, n = 30; GSE139384, n = 12; and GSE214417, n = 24; GSE185909, n = 35) were removed. The final analysis included four independent cohorts: GSE109887 (n = 78), GSE122063 (n = 100), GSE132903 (n = 195), and GSE243243 (n = 93). For pooled analyses, the four retained cohorts were merged after harmonizing gene features across studies. Genes with more than 50% missing values were removed, and remaining missing expression values were imputed using gene-wise median values, resulting in the finalized dataset Alzheimers_Disease_cleandata_final.csv used for pooled analyses before and after handling batch effect. In pooled analyses, cohort identity was treated as a batch variable to account for technical and study-specific effects, while brain region information was used to contextualize biological heterogeneity rather than as an explicit modeling covariate.

Differential Expression Analysis Within Cohorts: Differential gene expression analysis (DEA) was first conducted separately within each cohort to characterize cohort-specific Alzheimer's disease patterns. Analyses were performed using gene-wise linear models implemented in the `limma` package. For each gene g and sample i, normalized log-expression values were modeled as $y_{ig} = \beta_{0g} + \beta_{1g}\,\mathrm{AD}_i + \varepsilon_{ig}$, where AD_i indicates disease status [4]. Empirical Bayes variance moderation was applied to stabilize gene-wise variance estimates across the high-dimensional expression space. Statistical significance was assessed using moderated t-statistics with Benjamini–Hochberg false discovery rate (FDR) control.

Pooled Differential Expression Analysis and Batch Effects: To evaluate the impact of cross-cohort batch effects, gene expression data from all retained cohorts were pooled and analyzed both without and with cohort adjustment. Batch effects were addressed using a linear-model–based adjustment in `limma`, treating cohort identity as a batch variable while preserving disease-associated variation. Differential expression analysis was then repeated on the batch-corrected data. The batch-adjusted model can be expressed as $y_{ig} = \mu_g + \alpha_{b(i)g} + \beta_{1g}\,\mathrm{AD}_i + \varepsilon_{ig}$, where $\alpha_{b(i)g}$ denotes the

cohort-specific effect. Dimensionality reduction using UMAP was applied before and after batch correction to assess changes in cohort-driven structure.

Feature Selection Analysis:: To assess the impact of batch effects on downstream feature selection, we applied LASSO and random forest models to the data and compared the selected features before and after batch adjustment.

4 Results

Within-cohort differential expression analyses revealed substantial heterogeneity in Alzheimer's disease–associated transcriptional signals across cohorts (Figure 1). The volcano plots (Figure 1a–d) show the $\log_2$ fold change in gene expression between Alzheimer's disease and control samples plotted against the $-\log_{10}$ p-value. Genes highlighted in red meet both the statistical significance threshold and the fold-change criterion. The corresponding heatmaps of the top 20 differentially expressed genes (Figure 1e–h) further highlight cross-cohort heterogeneity. While cohort GSE122063 show clearer separation between Alzheimer's disease and control samples, others exhibit more overlapping expression patterns, indicating limited discriminative signal.

To assess the batch effects in pooled dataset, we visualized samples using UMAP colored by cohort labels (Figure 2). Prior to batch effect correction, samples clustered strongly by cohort, suggesting that batch structure dominated the low-dimensional representation (Figure 2a). After batch effect correction, samples from different cohorts exhibited substantially increased mixing in the UMAP space (Figure 2b), indicating effective attenuation of batch-driven variation. Together, these results demonstrate that cross-cohort heterogeneity substantially influences both differential expression results and global sample structure, motivating the need for cohort adjustment in pooled analyses.

LASSO selected 84 genes before batch adjustment and 116 genes after adjustment, with only 18 overlapping (Jaccard index = 0.099), indicating substantial sensitivity of feature selection to batch effects, which emphasizes the impact of cohort adjustment. Figure 2 comparison shows that feature importance rankings from the random forest model change substantially after batch correction, with most highlighted genes deviating from the diagonal (red dashed line), indicating that batch effects strongly influence feature prioritization and that only a limited subset of features retain similar importance before and after adjustment.

5 Conclusion

This study demonstrates that cross-cohort heterogeneity has a meaningful impact on gene–disease association analyses in multi-cohort Alzheimer's disease gene expression data. Differences between cohorts, including technical and experimental factors, substantially influence differential expression results and downstream feature selection, such that conclusions drawn from pooled analyses can change depending on whether batch effects are accounted for. While cohort adjustment reduced dominant batch-driven separation and revealed stronger disease-associated signals, residual clustering remained after correction, suggesting that not all variation can be attributed to removable technical effects. Such structure may reflect genuine biological differences across brain regions or study designs, as well as unmeasured technical factors.

Several limitations should be noted. Cohort sample sizes were modest, and larger cohorts contributed more strongly to pooled analyses, which is common in integrative studies. In addition, region- or laboratory-matched biological replicates were not available, limiting our ability to fully disentangle biological heterogeneity from technical variation. Future studies incorporating larger

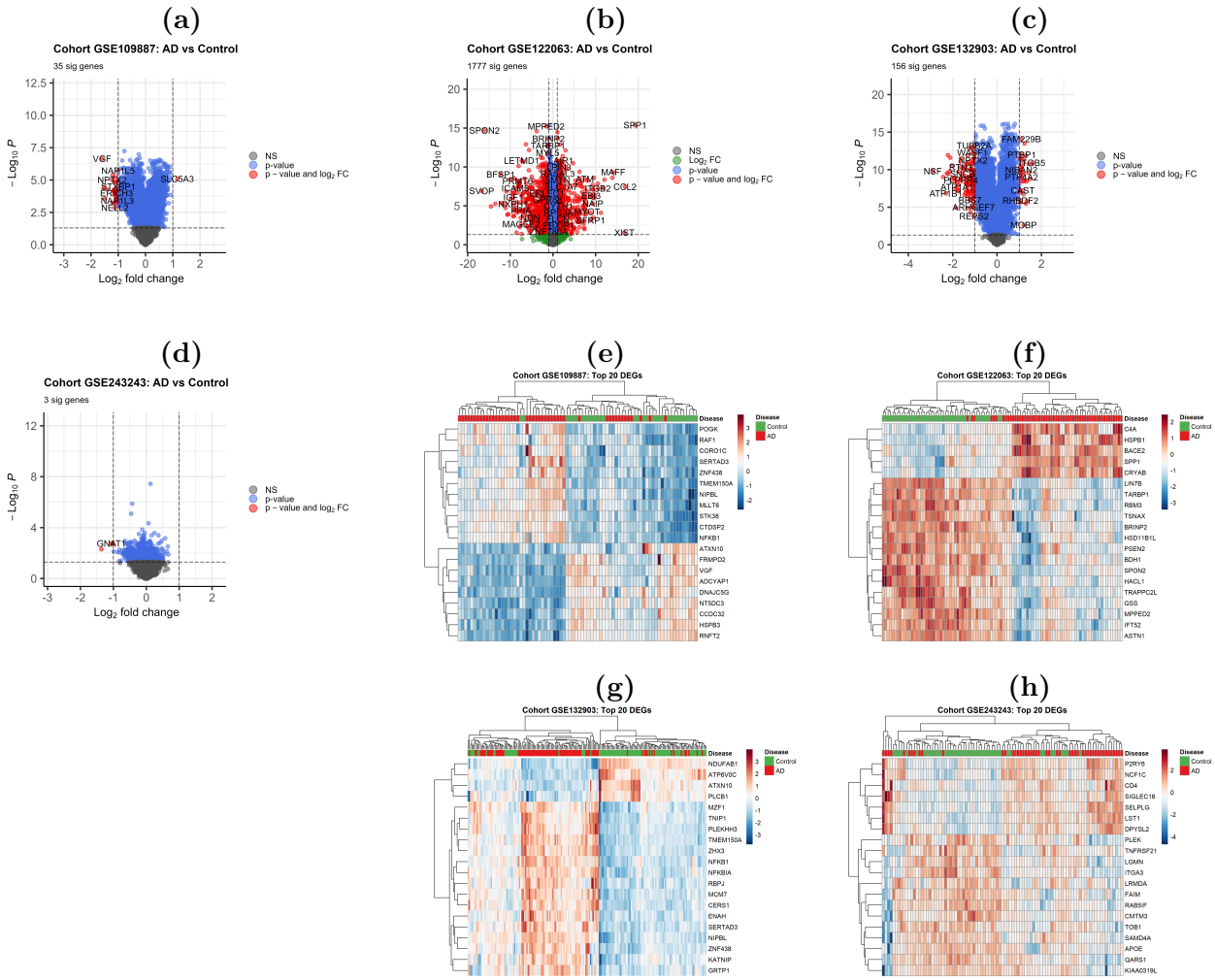

Figure 1: Within-cohort differential expression results. Panels (a)–(d) show volcano plots for AD vs control within individual cohorts: (a) GSE109887, (b) GSE122063, (c) GSE132903, and (d) GSE243243. Panels (e)–(h) show heatmaps of the top 20 differentially expressed genes for the corresponding cohorts: (e) GSE109887, (f) GSE122063, (g) GSE132903, and (h) GSE243243.

cohorts and replicated measurements across regions or laboratories would further strengthen robustness assessments.

Overall, these findings highlight the importance of batch-aware analysis and robustness evaluation when integrating heterogeneous transcriptomic datasets, and underscore that careful handling of cross-cohort heterogeneity is essential for reliable interpretation of disease-associated signals in Alzheimer's disease research.

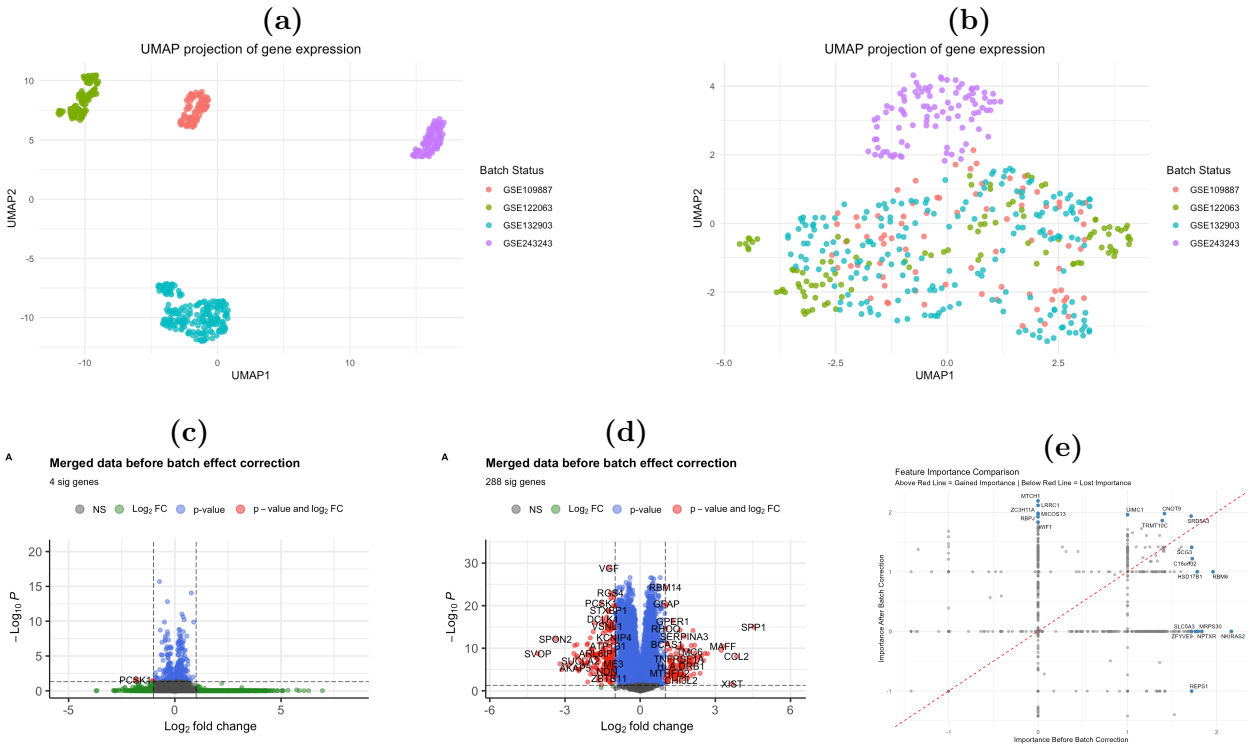

Figure 2: Pooled analysis results before and after cohort adjustment. (a) UMAP embedding before batch effect correction. (b) UMAP embedding after batch effect correction. (c) Differential expression analysis without cohort adjustment. (d) Differential expression analysis with cohort adjustment. (e) Comparison of random forest feature importance before and after batch correction.

6 References

1. Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Med. 2016 Nov 1;8(1):104. doi: 10.1186/s13073-016-0355-3. PMID: 27799057; PMCID: PMC5088659.

2. Collado-Torres, L., Burke, E. E., Peterson, A., Shin, J., Straub, R. E., Rajpurohit, A., Semick, S. A., Ulrich, W. S., BrainSeq Consortium, Price, A. J., Valencia, C., Tao, R., Deep-Soboslay, A., Hyde, T. M., Kleinman, J. E., Weinberger, D. R., & Jaffe, A. E. (2019). Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia. *Neuron*, *103*(2), 203–216.e8. https://doi.org/10.1016/j.neuron.2019.05.013

3. Leek, J., Scharpf, R., Bravo, H. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 733–739 (2010). https://doi.org/10.1038/nrg2825

4. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007. Epub 2015 Jan 20. PMID: 25605792; PMCID: PMC4402510.