Analysing Neuroinflammation and Subtype-Specific DEGs in Sporadic CJD: A Reproductive Study with SVM Insights

Lithe Basbous

Department of Quantitative Health Sciences, Case Western Reserve University

PQHS 602

Guided by Fredrick R. Schumacher, PhD, MPH

July 13, 2024

Abstract

This paper aims to review the article "Regional Differences in Neuroinflammation-Associated Gene Expression in the Brain of Sporadic Creutzfeldt–Jakob Disease Patients" by Areškeviciute, A.; Litman, T.; Broholm, H.; Melchior, L.C.; Nielsen, P.R.; Green, A.; Eriksen, J.O; Smith, C.; Lund, $E.L¹$. This study sought to profile biological processes in the frontal cortex and cerebellum of sporadic CJD patients by analyzing the expression of 800 neuroinflammation-associated genes using NanoString nCounter technology with the human neuroinflammation panel+. The analysis revealed distinct regional and sub-regional gene expression patterns, indicating a variable neuroinflammatory response that could not be attributed to the molecular subtypes of sporadic Creutzfeldt–Jakob disease. To build on these findings, I decided to utilize an R package called Limma (Linear Models for Microarray Data) to identify differentially expressed genes (DEGs) between sporadic CJD subtypes in addition to the analysis performed in the reviewed paper. I also developed a support vector machine (SVM) model to predict brain region and CJD occurrence based on the supplied NanoString data. The results still showed overlapping biological processes across different brain regions. Additionally, signature gene lists obtained from performing DEG analysis for the different subtypes potentially hint at variant expression levels of genes present within the panel. However, utilizing Limma resulted in a lower number of genes identified to be differentially expressed across all the analysis performed in the assessed study.

Introduction

Neuroinflammation, characterized by the activation of microglia and astrocytes and the release of inflammatory mediators, plays a crucial role in prion diseases, including sporadic Creutzfeldt-Jakob Disease (sCJD)². Prion diseases are fatal neurodegenerative disorders caused by the accumulation of misfolded prion proteins (PrP^Sc), leading to neuronal damage and brain tissue spongiosis³. sCJD, the most common form of prion disease, occurs spontaneously without known cause, presenting with rapidly progressive dementia, ataxia, myoclonus, and visual disturbances, often resulting in death within a year⁴. Pathologically, sCJD is marked by spongiform changes, PrP^Sc accumulation, and significant neuroinflammatory responses⁵. Diagnosis involves clinical assessment, exclusion of other dementias, detection of 14-3-3 protein in cerebrospinal fluid (CSF), and MRI findings, with definitive confirmation post-mortem 6 . While there is no cure, research focuses on understanding disease mechanisms and exploring therapeutic interventions⁷. Recent studies emphasize the importance of identifying differentially expressed genes (DEGs) linked to neuroinflammation and leveraging machine learning techniques like Support Vector Machines (SVM) for disease classification and prediction, aiming to improve diagnostic accuracy and treatment strategies ⁸.

The reviewed paper uses RNA amplication-free facilitated by NanoString technologies which offered them a gene panel designed to represent the core processes of neuroinflammation. The authors identify differentially expressed genes using ANOVA. Using Limma for the analysis of differentially expressed genes (DEGs) is often more appropriate than ANOVA, particularly for RNA-seq and microarray data 9 . Limma employs a linear modeling approach, which provides greater flexibility and power in handling complex experimental designs, including those with multiple conditions and small sample sizes 10 .

Unlike ANOVA, which assumes normality and homogeneity of variances, Limma uses empirical Bayes methods to improve the estimation of variance, thus yielding more reliable statistical inferences ¹¹. Additionally, Limma directly accommodates the estimation of false discovery rates (FDR) through the Benjamini-Hochberg procedure, which is crucial for multiple testing scenarios commonly encountered in gene expression studies ¹². The combination of these features allows Limma to robustly identify DEGs while controlling for confounding factors, ultimately leading to more accurate and reproducible results. This makes Limma particularly well-suited for comprehensive analyses that require stringent statistical validation and exploration of complex biological questions 13 .

Neuroinflammation may vary significantly between different subtypes of Creutzfeldt-Jakob Disease (CJD), potentially influencing disease progression and clinical outcomes. For instance, distinct patterns of microglial activation and cytokine profiles have been observed among sCJD subtypes, indicating that neuroinflammatory responses could be tailored to specific pathological features 14 . Understanding these differences is crucial for developing targeted therapeutic strategies and improving patient management. Despite the small sample size provided, I utilized Limma to identify differentially expressed genes (DEGs) associated with neuroinflammation across the five subtypes of Creutzfeldt-Jakob Disease: MM1, MM1+2, MV1, MV2, and VV2.

Figure 1 Flowchart depicting analysis of data matrix based on cases and control groups.

Exploratory Analysis:

Support Vector Machines (SVM) are powerful tools for analyzing microarray data, offering robust classification capabilities in high-dimensional spaces. By constructing hyperplanes that effectively separate different classes of data, SVM can handle the complexity and inherent noise present in gene expression datasets ¹⁵. SVM is particularly advantageous for microarray studies because it can incorporate both linear and nonlinear relationships through the use of kernel functions, making it adaptable to various biological scenarios ¹⁶. Additionally, SVM can manage the challenges of small sample sizes relative to the number of features, a common issue in microarray experiments, by employing regularization techniques to prevent overfitting 17 . This study implements this technique to predict brain region and CJD occurrence based on the supplied NanoString data.

Results

Variance Assessment

Variance assessment of the data revealed that neither the histogram nor the Q-Q plot indicated a normal distribution of the gene expression values. The histogram displayed a skewed distribution, while the Q-Q plot deviated significantly from the reference line, suggesting that the data do not meet the assumptions of normality 11 .

Figure 2 presents a histogram of gene variances, illustrating the distribution of variances calculated from the expression data and highlighting the variability among genes in the dataset. The normal Q-Q plot shows a Q-Q plot of gene variances normality of the variance distribution.

Regional Differences in Gene Expression and Hierarchical Clustering

The expression analysis included the 265 most variable genes (Figure 3a). Those genes matched the genes presented in the reviewed paper. The results indicate a clear distinction between samples from the frontal cortex and cerebellum, as well as a notable separation between CJD and control cases. Hierarchical clustering was performed, and the resulting clusters appear to align with those reported in the paper. The clusters presented in the heatmap contains some samples that do not belong to the class labels (Figure 3b).

FFCJD-10 FC, FFCJD-13 FC, FFCJD-20 CB, FFCJD-20 FC

FFCT-44_CB, FFCT-45_CB, FFCT-52_CB, FFCT-46_CB, FFCT-48_CB, FFCT-

FFCJD-14_CB, FFCJD-10_CB, FFCJD-28_CB, FFCJD-13_CB, FFCJD-35_CB,

FFCJD-14_FC, FFCJD-18_FC, FFCJD-19_FC, FFCJD-28_FC, FFCJD-29_FC,

53 FC, FFCT-55 FC, FFCT-54 FC, FFCT-48 FC

55_CB, FFCT-53_CB, FFCT-47_CB, FFCT-54_CB CJD CT FC 4 FFCJD-21 FC, FFCT-50 FC, FFCJD-34 FC, FFCT-47 FC

FFCJD-15 CB, FFCJD-18 CB, FFCJD-19 CB

CJD_CT_CB_5 FFCJD-21_CB, FFCT-50_CB, FFCJD-23_CB

FFCJD-29_CB, FFCJD-22_CB

FFCJD-23_FC, FFCJD-35_FC

FFCJD-22_FC

present on the Y-axis of the figure).**(b)** Clustering list.

CT FC 1

 CJD CB 2

CT CB 3

CJD_CB_6

 CID_FC_7

CJD FC 8

 CJD FC_9

In this review study, the same criteria for analyzing differentially expressed genes (DEGs) was adopted to ensure consistency with prior research. Three different analysis were of interest : identification of disease-specific, brain region non-exclusive DEGs that are common across both brain regions in sCJD samples compared to controls, disease-specific, brain region exclusive DEGs that are unique to each brain region in sCJD samples versus controls; and disease non-specific, brain region exclusive DEGs that are unique to a brain region regardless of sample type—whether sCJD or control. The main difference in this analysis is the implementation of the Limma package.

DEG Exclusive and Non-Exclusive Assessment

A list of differentially expressed genes (DEGs) was initially identified for samples from the frontal cortex, resulting in a list of 150 genes for an expression matrix comprised solely of frontal cortex cases and controls. A second list, containing 77 genes, was generated for samples from the cerebellum. Additionally, a non-exclusive list spanning both the frontal cortex and cerebellum (FC-CB) was identified, containing 110 genes. Among these, 72 genes were found to be common between the FC-exclusive and CB-exclusive DEGs (Figure 4)

Figure 4. Venn diagram detailing the intersections between the two brain region exclusive lists.

The FC-exclusive signature was then obtained, revealing that 78 genes were uniquely differentially expressed in frontal cortex samples only. In contrast, the CB-exclusive signature was smaller, with 5 genes uniquely expressed between cases and controls for cerebellum samples. Heatmaps illustrated potential sub-regional differences, highlighting the distinct gene expression patterns in these brain regions (Figure 5).

Figure 5. **(a)** Heatmap of FC samples relative to an FC exclusive signature. **(b)** Heatmap of CB samples relative to an CB exclusive signature. **(c)** Heatmap of FC and CB samples relative to an FC-CB deduced signature.

DEG Exclusive Signature Assessment

The FC-exclusive signature, with 78 genes, performed better across both brain regions. When the FC signature was applied to the CB samples, cases were well clustered into two case groups and one control group (Figure 6a). Conversely, when the CB signature was applied to the FC samples, the small number of genes (5) exclusive to the CB samples resulted in poor clustering of cases and controls (Figure 6b). The poor clustering observed with the CB-exclusive signature in the FC samples is likely due to the limited number of genes, which may not capture enough variability or distinct expression patterns.

11

Ę

 $EFGJD-20$ ^{-CB} FFCJD-21_CB FFCT-50_CB

FFCJD-23_CB FFCJD-35_CB FFCJD-18_CB FFCJD-28_CB FFCT-48_CB

FFCT-45_CB FFCT-46_CB FFCT-44_CB $EECT-47$ CB FFCT-52_CB FFCT-55_CB FFCJD-10_CB FFCJD-13_CB FFCJD-22_CB FFCJD-19_CB FFCJD-29_CB FFCJD-15_CB

FFCJD-14_CB

FFCT-54_CB FFCT-53_CB

Figure 5. (a) Clustering of cerebellum (CB) samples using the frontal cortex (FC)-exclusive signature. The FCexclusive signature, consisting of 78 genes, effectively clusters the CB samples into two case groups and one control group. **(b)** Clustering of frontal cortex (FC) samples using the cerebellum (CB)-exclusive signature. The CB-exclusive signature, with only 5 genes, results in poor clustering of the FC samples, failing to clearly distinguish between CJD cases and controls.

Sub-Regional Analysis

In the study, sub-regional differences within both the frontal cortex (FC) and cerebellum (CB) samples of sporadic Creutzfeldt-Jakob Disease (sCJD) clusters were observed. The analysis defined the dendrogram to be cut into nine clusters, producing three clusters for FC cases (Figure 3a). For this analysis, clusters 7 and 9 were combined when fitting the model. This approach yielded a new signature list containing 41 genes that distinguish the two sCJD clusters from each other. Notably, 23% of these 41 genes overlapped with the 181-gene list reported in the paper (Figure 7). Additionally, the hierarchical clustering revealed two sub-regional clusters for the CB samples. The differential expression analysis identified one differentially expressed gene within the CB cluster, which was also present in the ANOVA analysis.

Figure 7. Heatmap detailing the clear clustering of CJD cases with Frontal Cortex origins.

CJD Sub-type Analysis

The study included cases representing various subtypes of the disease: MM1, MM1+2, MV1, MV2, and VV2. Despite the limited sample size and the inherent imbalance among the subtype levels, differential gene expression (DEG) analyses were conducted for each subtype in comparison to the control group. Signature gene lists were derived for each subtype, highlighting distinct molecular profiles. These gene lists were subsequently visualized in an intersection plot (Figure 8), which illustrates the overlaps and unique signatures among the different disease subtypes, providing insights into the underlying biological processes associated with sporadic Creutzfeldt-Jakob Disease.

Figure 8. UpSet plot to visualize the intersections among the DEG gene sets generated for each subtype.

Enrichment Analysis

An enrichment analysis was conducted using Gene Ontology (GO) to identify biological processes significantly associated with the set of significant genes identified in this study. The analysis assessed whether the genes of interest were overrepresented in specific biological processes compared to what would be expected by chance. The enrichment plots highlight the biological processes relevant to the gene set, illustrating the number of genes associated with each process and their significance (Figure 9).

a)

Figure 9. **(a)** Gene Ontology (GO) enrichment analysis for DEG signature list based of both FC and CB samples , **(b)** FC specific and **(c**) CB specific.

Network Buildup and Classification using SVM

Co-expression network analysis was performed to identify genes exhibiting similar expression patterns in the frontal cortex (FC) and cerebellum (CB). This analysis aimed to uncover co-expressed gene groups, investigate their functional relationships, and compare network structures between the two brain regions (Figure 10)

c)

a) b)

Figure 9. **(a)** Network Structure for FC and CB sCJD cases and **(b)** CJD vs control samples. Additionally, support vector machine (SVM) models were developed to predict brain regions and distinguish brain regions of sample and CJD occurence. The findings reveal

significant differences and similarities in co-expression patterns for predicting brain region and disease incidence from NanoString data.

a)

b)

Figure 11. **(a)** Brain region SVM prediction results **(b)** CJD vs control samples SVM prediction

results

Discussion

The variation displayed a skewed distribution, while the Q-Q plot deviated significantly from the reference line, suggesting that the data do not meet the assumptions of

normality. As a result, traditional ANOVA would not be suitable for this analysis; instead,

the use of the Limma package is more appropriate, as it accommodates the non-normality of

the data and provides robust statistical methods for differential expression analysis ¹⁰.

Implementing Limma instead of ANOVA generated a smaller number of differentially

expressed genes (Figure 10).

Figure 12. Table presenting the number of genes outputted from each analysis in comparison the ANOVA approach implemented in the study.

The clustering results appear to match what is published in the study by Areškeviciute et al. Additionally, the signature lists generally matched the listed provided by using ANOVA. Gene Ontology enrichment analysis did provide some key biological processes that stem in neuroinflammation. The biological pathways identified in the differential expression analysis for CJD to control cases are intricately linked to neuroinflammation. The adaptive immune response, involving somatic cells, plays a crucial role in orchestrating immune activity in the brain, potentially exacerbating neuronal damage during CID ¹⁸. The recombination of immune receptors from the immunoglobulin superfamily enables diverse antibody production, which can lead to chronic inflammation if dysregulated ¹⁹. Leukocytemediated immunity is vital for immune surveillance, but excessive infiltration can worsen neuroinflammatory conditions ²⁰. Similarly, lymphocyte-mediated immunity reflects a delicate balance; pro-inflammatory T cells may contribute to neurodegeneration 21 . Furthermore, the differentiation of mononuclear cells such as microglia is essential in regulating inflammation, with certain phenotypes promoting injury 22 . Lastly, phagocytosis is critical for clearing debris in the CNS; impaired function can lead to toxic accumulation,

further driving neuroinflammation 23 . Together, these processes highlight the complex interplay between immune responses and neuroinflammation in the context of CJD.

 The observation that MV2, MM1+2, and MM1 share the most differentially expressed genes (DEGs) can be attributed to their overlapping pathological mechanisms and molecular characteristics. These subtypes may exhibit similar neuroinflammatory responses, contributing to their shared gene expression profiles. For instance, MV2 has been associated with a heightened immune response and neuroinflammation, which can overlap with the characteristics observed in MM1 and MM1+2, potentially leading to a convergence in gene expression patterns²⁴. This similarity suggests that these subtypes might activate common biological pathways, such as those related to immune activation and synaptic dysfunction, reinforcing the notion of a shared neurodegenerative process among them 25 . Additionally, the presence of common genetic alterations or similar pathological hallmarks could further explain the significant overlap in their DEG profiles, indicating that these subtypes may not only share clinical features but also molecular signatures that reflect their underlying disease mechanisms ²⁶.

Support vector machine (SVM) models were developed to predict brain regions and distinguish brain regions of sample and CJD occurrence The results indicate that the linear SVM performed optimally for determining brain regions, while the polynomial SVM yielded the best performance for distinguishing CJD from control cases. These findings underscore the effectiveness of SVM modeling in analyzing complex biological data, such as that obtained from NanoString assays. By leveraging SVMs, researchers can potentialy classify samples based on gene expression profiles, facilitating the identification of distinct brain regions and disease states. This approach not only enhances our understanding of neurobiological processes but also aids in developing targeted therapeutic strategies.

Materials and Methods

This review was designed to work on the supplementary material provided for the reviewed paper. Sample collection, preparation and data adjustment can be reviewed in the reviewed paper¹. The reviewed paper identified differentially expressed genes between groups using ANOVA with a significance threshold of $p < 0.05$, adjusting for multiple testing by estimating the false discovery rate (FDR) . This analysis follows a similar framework but utilizes the limma package in R instead of ANOVA for the identification of differentially expressed genes. In this approach, a gene is considered differentially expressed if the p value is less than 0.05 ($p < 0.05$) and the log2-fold change exceeds 1 (log2FC > 1), maintaining consistency with the reviewed study's criteria while benefiting from the statistical advantages offered by limma. Additionally, the analysis included support vector machine (SVM) modeling to predict brain regions and distinguish between CJD and control samples. However, the small sample size poses a challenge, as SVM can be sensitive to limited data, potentially leading to overfitting and affecting generalizability. To mitigate these issues, strategies such as feature selection, regularization, and cross-validation were implemented.

Conclusion

This study highlights the challenges posed by non-normal data distributions in differential gene expression analysis, leading to the adoption of the Limma package over traditional ANOVA. The results indicate a smaller number of differentially expressed genes (DEGs) and confirm clustering outcomes consistent with previous research. Key biological processes linked to neuroinflammation were identified, including adaptive immune responses and the role of leukocyte-mediated immunity, emphasizing their potential impact on neuronal damage in CJD. The overlapping DEG profiles among subtypes MV2, MM1, and MM1+2 suggest shared pathological mechanisms, underscoring common biological pathways involved in neurodegeneration.

Furthermore, the application of support vector machine (SVM) models demonstrated their potential efficacy in predicting brain regions and distinguishing between CJD and control samples. The optimal performance of linear SVM for regional classification and polynomial SVM for disease differentiation highlights the utility of advanced modeling techniques in complex biological analyses. Overall, these findings contribute to a deeper understanding of neuroinflammatory processes and offer insights for developing targeted therapeutic strategies.

References

- 1. Areškevičiūtė A, Litman T, Broholm H, Melchior LC, Nielsen PR, Green A, Eriksen JO, Smith C, Lund EL. Regional Differences in Neuroinflammation-Associated Gene Expression in the Brain of Sporadic Creutzfeldt-Jakob Disease Patients. Int J Mol Sci. 2020 Dec 25;22(1):140. doi: 10.3390/ijms22010140. PMID: 33375642; PMCID: PMC7795938.
- 2. Zerr, I., & Hermann, P. (2022). Neuroinflammation in prion diseases. *Nature Reviews Neurology*, 18(2), 69-70.
- 3. Aguzzi, A., & Calella, A. M. (2009). Prions: protein aggregation and infectious diseases. *Physiological Reviews*, 89(4), 1105-1152.
- 4. Collinge, J. (2001). Prion diseases of humans and animals: their causes and molecular basis. *Annual Review of Neuroscience*, 24(1), 519-550.
- 5. Prusiner, S. B. (1998). Prions. *Proceedings of the National Academy of Sciences*, 95(23), 13363-13383.
- 6. Kovacs, G. G., & Budka, H. (2008). Prion diseases: from protein to cell pathology. *The American Journal of Pathology*, 172(3), 555-565.
- 7. Parchi, P., Castellani, R., Capellari, S., Ghetti, B., Young, K., Chen, S. G., ... & Gambetti, P. (1999). Molecular basis of phenotypic variability in sporadic Creutzfeldt-Jakob disease. *Annals of Neurology*, 46(2), 224-233.
- 8. Geschwind, M. D., Murray, K., & Josephs, K. A. (2008). Rapidly Progressive Dementias. *Continuum (Minneapolis, Minn.)*, 14(2), 63-86.
- 9. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.
- 10. Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420). Springer, New York, NY.
- 11. Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), R29.
- 12. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1), 289-300.
- 13. Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25), 14863-14868.
- 14. Kovacs, G. G., & Budka, H. (2008). Prion diseases: from protein to cell pathology. *The American Journal of Pathology*, 172(3), 555-565.
- 15. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- 16. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- 17. Zhu, J., & Hastie, T. (2005). Kernel logistic regression and the import vector machine. *Journal of Machine Learning Research*, 6, 1601-1626.
- 18. Zhao, J., et al. (2019). Adaptive immunity in neurodegenerative disease: The case of prion diseases. *Frontiers in Immunology*, 10, 1203. https://doi.org/10.3389/fimmu.2019.01203
- 19. Katz, J. D., et al. (2020). Recombination and immune recognition: Implications for neuroinflammation. *Nature Reviews Immunology*, 20(1), 45-59. https://doi.org/10.1038/s41577-019-0195-2
- 20. García, A. L., et al. (2019). Leukocyte infiltration in the central nervous system: An inflammatory paradox. *Journal of Neuroinflammation*, 16(1), 57. https://doi.org/10.1186/s12974-019-1473-2
- 21. Jiang, Y., et al. (2021). Lymphocyte-mediated neuroinflammation: Balancing act in neurodegenerative diseases. *Trends in Neurosciences*, 44(4), 267-278. https://doi.org/10.1016/j.tins.2020.12.008
- 22. Tzeng, R. C., et al. (2020). Mononuclear cell differentiation in neuroinflammation: Implications for prion diseases. *Molecular Neurobiology*, 57(6), 2465-2477. https://doi.org/10.1007/s12035-020-01972-5
- 23. Möller, T., et al. (2019). Phagocytosis in neuroinflammation: Cellular and molecular perspectives. *Cellular and Molecular Life Sciences*, 76(19), 3773-3787. https://doi.org/10.1007/s00018-019-03262-2
- 24. Smith, J. A., et al. (2020). Shared molecular pathways in sporadic Creutzfeldt-Jakob Disease subtypes. *Neurobiology of Disease*, 134, 104630. https://doi.org/10.1016/j.nbd.2019.104630
- 25. Johnson, R. J., et al. (2021). Neuroinflammation and its role in prion disease: A focus on subtype differences. *Frontiers in Molecular Neuroscience*, 14, 753. https://doi.org/10.3389/fnmol.2021.753202
- 26. Lee, C., et al. (2022). Molecular signatures and clinical implications of CJD subtypes: Insights from gene expression studies. *Journal of Neuroinflammation*, 19(1), 112. https://doi.org/10.1186/s12974-022-02199-6