Term Project Report

3/18/2019

STAT 465, Dr. Wang

Will Minor, Jide Anene, Sara Wixon

Using Hierarchical Clustering to Characterize Gene Expression In Alzheimer's Disease and Chronic Traumatic Encephalopathy

Project Contributions:

Final Report:

Introduction - Sara

Background- Sara and Will

Project Methods- Jide and Will

Results- Jide, Will, Sara

Discussion - Sara and Jide

Conclusion: Sara

Final Presentation - Sara, Will, and Jide

Code:

Data Processing - Jide

Clustering Analysis - Will, Sara

Overall, we all contributed near equally to this project. Everyone contributed in different ways throughout the process and each person's contributions were necessary for the successful completion of this project.

Executive Summary

Neurodegenerative diseases can be very devastating. However, not much is known about the disease mechanisms, and there is a dearth of effective treatments. A relatively new neurodegenerative disease, chronic traumatic encephalopathy (CTE), has recently been linked to both football and Alzheimer's disease (AD).[1] Playing football is a risk factor for both CTE and AD due to the repetitive head trauma that is characteristic of the sport.[2] CTE and AD share a controversial relationship, but some evidence indicates that CTE may promote the development of AD. Post-mortem biopsies of both AD and CTE patients share some key pathological features, such as the overexpression of hyperphosphorylated tau protein.[3] Tau protein has not been proven to cause either disease, but additional shared pathological features that may be important in identifying a relationship.

Significance. Determining the shared features of CTE and AD is important for a number of reasons. First, identifying shared features, such as gene expression patterns, may be useful for defining mechanisms and pathological effects. The determination of disease mechanism would yield new drug targets--and currently there are no effective treatments for either AD or CTE.[1,4] Second, identifying disparate features in those same expression patterns could help distinguish between the two diseases. Currently, clinical presentation of the two diseases is very similar, and only medical history helps decide which is the more likely cause. As treatments improve, it is imperative to have an accurate diagnosis so the chosen treatment is as effective as it can be.

Specific aims.

1. Determine significant differences in gene expression between conditions
2. Determine patterns over time in gene expression in each model
3. Compare patterns and gene expression in the two models

Background

Alzheimer's Disease

Alzheimer's disease (AD) is the most common cause of dementia in the U.S. and is the sixth greatest cause of death in the U.S.[7] The disease is characterized by the accumulation of amyloid proteins as well as tau tangles in neurons. Amyloid beta oligomers (ABOs) are widely believed to be the primary cause of AD. AD is instigated by a large accumulation of ABOs which then start to selectively bind to neurons.[8] Once bound, ABOs initiate a host of cellular signal cascades that lead to neuron cell death. The microtubule binding protein known as tau, becomes abnormally phosphorylated in AD, which causes it to form aggregates within neuron cells. Tau tangles, as these aggregates are called, are believed to be a main cause of cell death and are correlated with ABO binding to neuron cells.[9] As AD progresses, more neurons progressively die, leading to dementia and death.

CTE

CTE is a neurodegenerative disorder associated with repetitive traumatic brain injury.[1] The symptoms are similar to those of other neurodegenerative diseases, and include impairment of memory, motor function, decision making, and behavioral changes, but may also include mental health changes, including suicide.[1,4] While the pathology of CTE is still unknown, it is thought to be caused by repetitive head trauma, and is associated with athletes that play contact sports such as football or boxing.[4] There is currently no treatment or diagnostic for CTE, and prevention efforts are focused on preventing concussion by wearing helmets. Post-mortem biopsies have revealed that CTE is associated with reduced gray matter in the frontal and temporal lobes, as well as an anti-tau immune response, including neurofibrillary tangles, neuropil threads, dot-like grains, and astrocytic tangles.[1]

Project Methods

Overview

We have leveraged two sets of microarray data, hippocampal gene expression changes across different brain injury models (fluid percussion vs controlled cortical) in rats,[6] and hippocampal gene/protein expression and cognitive function in rats across the adult lifespan.[5] The rat brain injury data set was produced from Affymetrix Rat Genome U34 Array, and the rat lifespan dataset was produced from Affymetrix Rat Expression 230A Array. The differences in the arrays, as well as batch effects, made quantitatively comparing the two arrays difficult. Instead, we quantified significant gene expression individually, and then qualitatively compared the results between arrays.

Datasets. We chose a cognitive decline/aging dataset to serve as our AD model. This dataset used a water maze to track cognitive decline, analyzing hippocampal tissue at 3, 6, 9, 12, and 23 months in a *Rattus norvegicus* model. To model CTE, we chose another *Rattus norvegicus* model that used fluid percussion to induce moderate brain injury. Fluid percussion involves implanting a saline reservoir connected to a transducer. At the signal, set amounts of saline are rapidly injected into the surrounding neural tissue, causing damage.[6] Damage was assessed both by scoring of motor function and after death, by histological examination of the neural tissue.[6] Samples were taken 30 minutes, 4, 8, 24, 72, hours, and 21 days after the injury.

Objective. Our goal was to quantify the similarity between the molecular mechanisms that underlying brain damage after traumatic injury to the mechanisms that underlie cognitive decline during the natural process of aging. To this end, we performed a pathway analysis, also called a gene-set analysis (GSA), to identify sets of genes that are jointly associated, with the aim of gaining a comprehensive representation of the state of the brain cell throughout its response to brain trauma and the aging process. We used the classical cluster analysis method first introduced by Eisen et al (1998).[10]

Data Processing

For the CTE microarray dataset we used a method similar to Natale et al (2003)[6] to control for the effects of the surgical procedure of implanting the fluid percussion device into the rats cortical region. This means that for each probe set, we first normalized probe intensities from the injured rats (n = 12) to the mean signal intensity generated from the same cortical region from three naive rats. The naive rats in this context are control rats that were not surgically operated on. To be precise, for each probe-pair in the microarray dataset that represents the injured rats we took the difference between perfect-match (PM) probe-pair intensity and the average PM intensity for the microarrays that represent the naive rats. Probe set normalization was not applied to the AD microarrays.

We then calculated the discriminant score for the i^{th} probe-pair as $R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$, and for each probe-pair we performed the hypothesis $H_0 : median(R_i) = \tau$, corresponding to absence of transcript, and $H_A : median(R_i) > \tau$, corresponding to presence of transcript. The Wilcoxon signed-rank test was used to calculate a p-value for each probe-pair with τ =0.015. [15, 14, 13] A transcript was labeled “present” if the p-value associated with it was less-than 0.04, and a transcript was labeled “marginal” if the p-value associated with it was within the interval [0.04, 0.06]; we only considered transcripts that were rated “present.” To filter out genes that were not significantly expressed over time we generated multiple datasets corresponding to microarrays whose probe-sets were rated “present” in at-least 40%, 45%, 50%, and 55% for the CTE and AD microarrays, respectively.

To get expression intensity reads from the probe level AD and CTE microarray datasets we used the RMA (Robust Multichip Average) expression measure described in Irizarry et al (2003c).[18] Then we applied differential expression to filter out genes that were constant over time. To this end we used the two-sample Welch t-statistics to calculate p-values. Since conservative tests eliminate more genes than less-conservative tests, we defaulted to using the equal variance assumption, and the Bonferroni FWER adjustment [23] to calculate the p-values. Table 1 shows the number of genes filtered after each step in our two-step gene filtering process.

Table 1: Number of Genes after Filtering

Dataset	Num. of Genes	40% Filter	45% Filter	50% Filter	55% Filter
AD	**15923**	**10758 (17)***	**10625 (17)**	**10528 (17)**	**10432 (17)**
CTE	**8799**	**2753 (582)**	**2382 (488)**	**2030 (399)**	**1701 (340)**

*The value in the parenthesis indicates the number of genes left after filtering genes whose expression intensity measures were not significantly expressed over time)

Hierarchical Clustering and Heatmap Analysis

This method arranges genes according to the similarity of their gene expression patterns and is known to robustly associate genes of similar biological function.[10] This model uses an upper diagonal similarity matrix based on correlation distance to assign initial clusters to each gene in the dataset. Then each cluster is iteratively linked based on a similarity metric. For these times series, we decided to use correlation distance for the clustering, because we were most interested in trend similarity rather than value similarity, which would’ve been better suited for manhattan distance. Hierarchical clustering is an ideal analysis method for this project because it provides gene expression pattern information on clusters of similarly functioning genes, which allows for clear connections to biological function. For instance, if a disease involves the degradation of a particular protein, some clusters of regulatory proteins that behave abnormally, will have distinct patterns compared to their normal functioning cluster mates. This patterns can be clearly correlated to biological function and can allow for higher level understanding than simply associations between genes.

The clustering analysis model that we developed analyzed two cleaned, processed files that contained genes of significant expression that change over time for each disease. From **Table 1** we see that regardless of the filtering threshold for AD, 17 genes were identified using our 2-step filtering procedure, so in some respect the 45% threshold used in our clustering analysis is quite arbitrary. From **Table 1** we also see that AD and CTE datasets respond differently to our two-step filtering procedure. The number of differentially expressed genes in the CTE datasets is quite elastic to threshold choice, while for AD the number of differentially expressed genes is inelastic. This difference could be a consequence of the surgery used to implant CTE rats with the fluid percussion device activating more genes, or it could simply reflect that brain trauma activates more genes as a response relative to the aging process. The reasons for these differences are confounding and beyond the scope of this paper, but could inspire further research. To get a manageable number of genes for CTE we chose the 55% threshold which identified 340 genes of interest. Note that if we chose a lower threshold number there would've been a large number of genes to a cluster, which wouldn't have likely provided significant temporal trends due to clusters containing multiple gene types (ie. regulatory genes, structural genes, etc.). Overall, with added complexity the subsequent results would be harder to interpret. The txt file outputs of the data processing code were then time labeled and sorted in csv files.

The clustering analysis code then took the csv files and converted them from data frames into numerical matrices. These matrices were subsequently standardized and then initial dendrograms were made. Then, the data sets were hierarchically clustered and each cluster was displayed as a heatmap. The ideal number of clusters for AD was three, so that all genes appeared in all heatmaps. For CTE, the ideal number of clusters was 10 using similar rationale. The characterization patterns for the heatmaps were based off of the model in Kadish et al. (2009)[5], the AD data set paper, which is seen in **Figure 1**.[5] These eight expression patterns characterize typical regulation patterns for genes over time courses. As seen in **Figure 3**, the heatmap outputs gene expression that can either be upregulated (in red), which signifies increased gene expression, or downregulated (in green) which indicates decreased gene expression. Additionally, there is little to no effect on gene expression (in black) such that the gene is expressed at the same level over time. After the genes were clustered, the gene probe names were input into the Affymetrix database where they were translated into official gene names. Additionally, information on gene function was given alongside the gene names. The analysis of these particular genes will be the basis for the discussion section. Overall, hierarchical clustering heatmaps provide visual representations of gene expression patterns over time, which allows conclusions to be drawn on the biological significance of the gene expression under the specified conditions.

Results

Clustering Analysis

The results of the hierarchical clustering analysis are the dendrograms, heatmaps, and expression regimes for each gene cluster. In **Figure 2**, both the dendrogram for AD (**2a**) and CTE (**2b**) are shown. **Figures 3a-c** show AD heatmaps while **Figures 4a-c** show CTE heatmaps. A qualitative analysis of the heatmaps follows.

AD. Because our method of identifying significant genes was rather conservative, only 17 genes were determined to be significant in the AD model. We characterized three expression patterns: **Figure 3a** shows a late down expression pattern, where gene expression decreased late in life. **Figure 3b** shows a late up pattern, where gene expression increased late in life. Finally, **Figure 3c** shows a pattern that may be likened to a cycle. Cluster 1 mainly consisted of immune modulators and lysosomal storage/transport genes. This is interesting because neurodegenerative diseases are known to share features with lysosomal storage diseases, such as Niemann-Pick disease.[11] Cluster 2 contained mostly antigen presentation genes and a few others of interest. In particular, Cd74, a beta-amyloid binding protein that is implicated in AD, and Npc2, a gene associated with a lipid storage disorder that is characterized by progressive neurodegeneration.[12] Both of these genes showed increased expression over time. Cluster 3 only contained a phospholipid binding protein and a macrophage 1 protein.

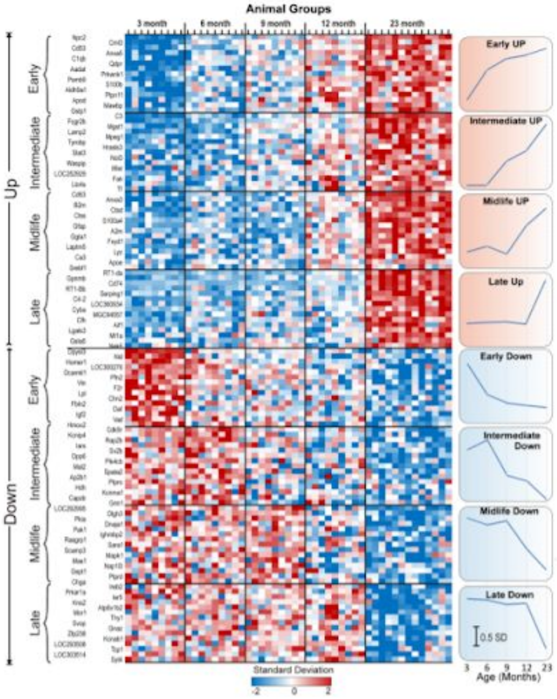

Figure 1: Heat Map of Age Related Decline in Kadish Et Al. The 8 expression patterns were utilized for gene expression pattern characterization in this project.

CTE. The CTE heatmaps seen in **Figure 4a-c** are three heatmaps out of 10 total, and have the most clear expression pattern. The other seven will be present in the Appendix under **Figure A**. The expression pattern for Cluster 1, Cluster 3, and Cluster 8, which are present in **Figure 4 a-c**, are late up, early down, and upregulation spike at 8hr, respectively. Cluster 1 has a wide variety of genes of various functions. Some genes of particular interest are Sparc, which is involved in ossification, Htr3a, a serotonin receptor involved in anxiety, depression, & PTSD, and lastly, some genes involved in memory and locomotion. Calcium release and ossification are associated with inflammation and injury, which is consistent with the injury the animal model suffered. CTE is also known to be associated with mental health problems such as PTSD, anxiety, depression, and suicide,[1,4] so the change in Htr3 could be significant. Cluster 3 has a wide variety of genes, that are mainly regulatory in nature. Some notable genes include: a few kinases, including Cdk17, Fn3k, and Dgk. Additionally, Ptn, which is involved in ossification and brain development, Cited2, a transcription activation factor, and Alcam, a cell adhesion gene, also exhibit decreased expression early in the injury. Cluster 8 had lots of genes associated with oxidative stress, wound healing, prostaglandin synthesis, and ossification. Two genes of interest are Fcgr2a (Fc of IgG antibodies) and Cd74, which is the beta-amyloid binding protein gene that was also found in the AD datasets. The other seven clusters did not have defined expression patterns.

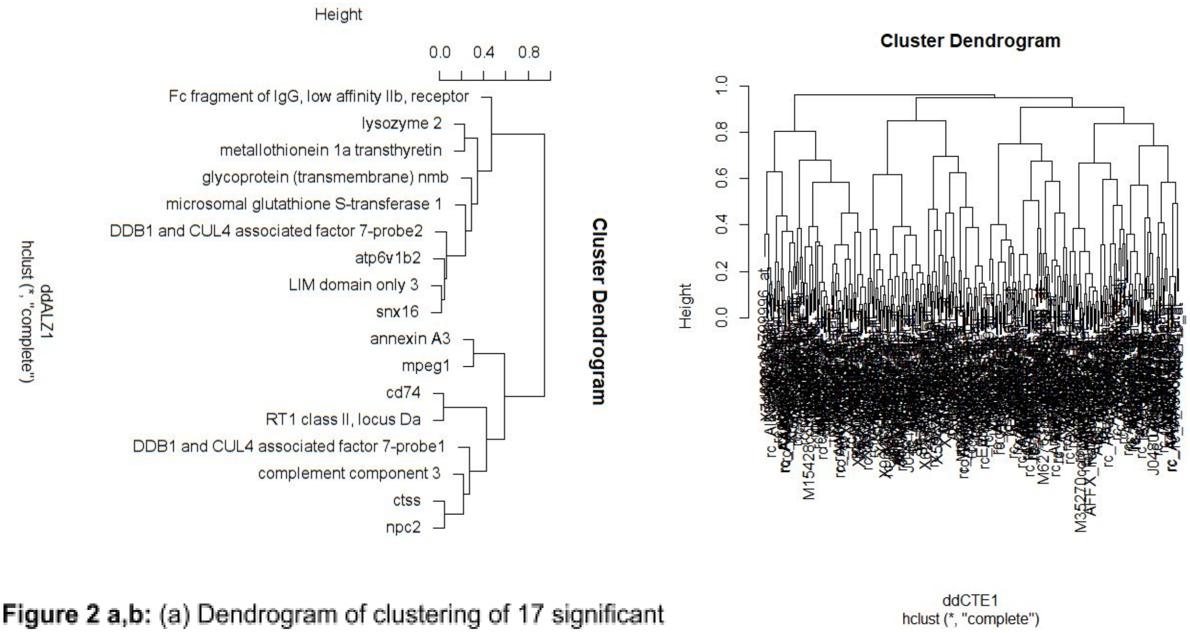

Figure 2 a,b: (a) Dendrogram of clustering of 17 significant Alzheimer's Disease Genes at a 45% threshold. (b) Dendrogram of clustering of 340 significant CTE genes at a 55% threshold.

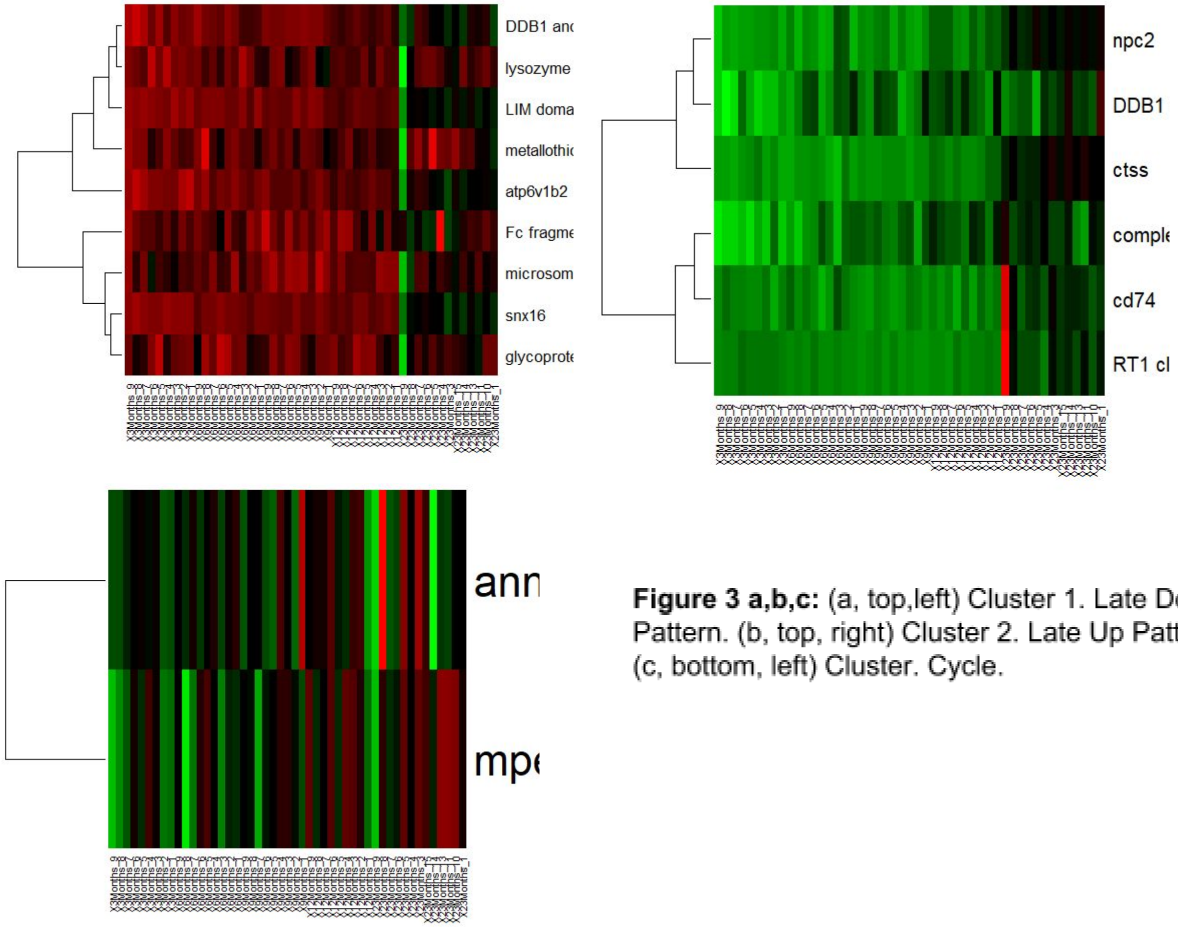

Figure 3 a,b,c: (a, top,left) Cluster 1. Late Down Pattern. (b, top, right) Cluster 2. Late Up Pattern. (c, bottom, left) Cluster. Cycle.

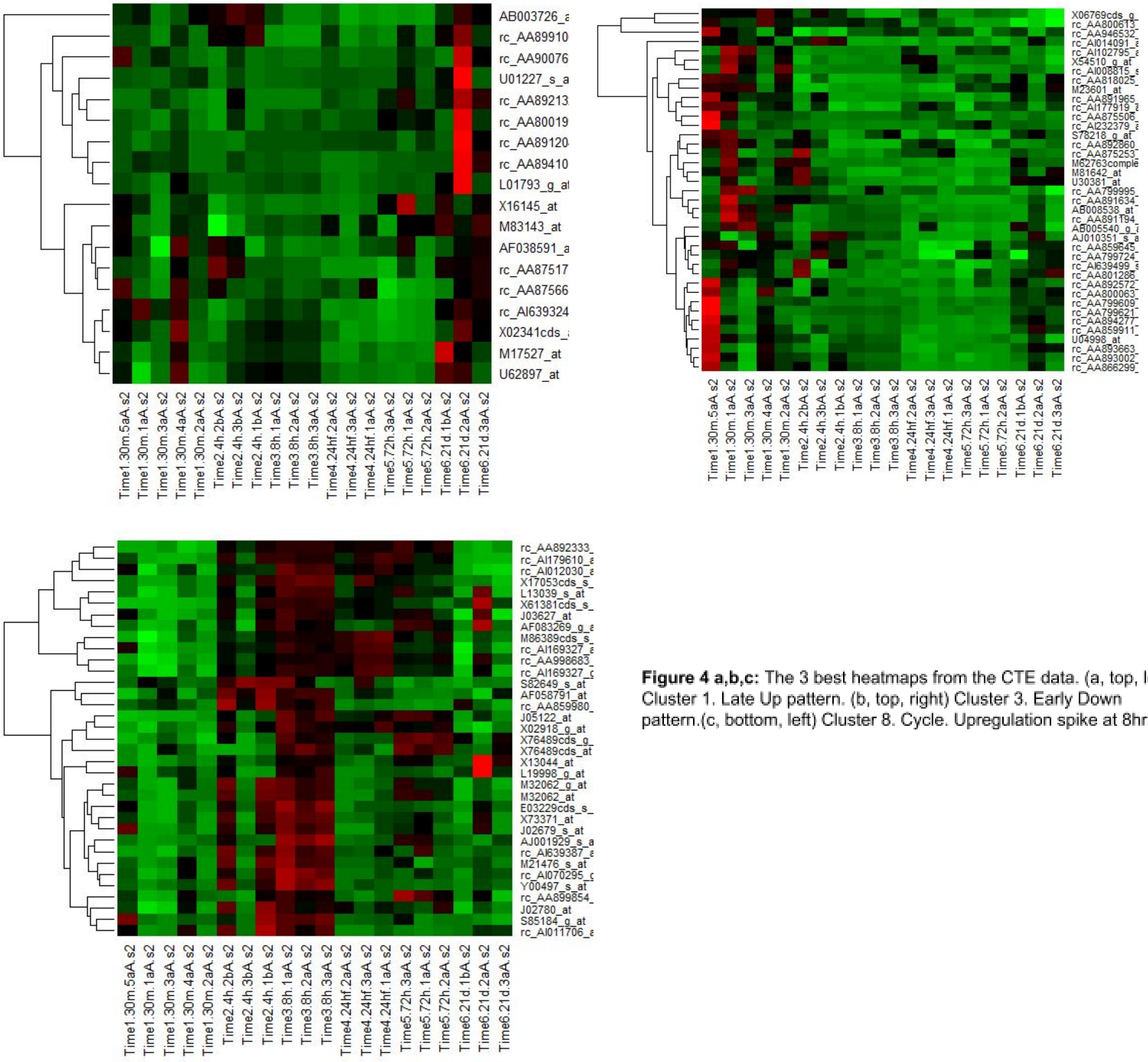

Figure 4 a,b,c: The 3 best heatmaps from the CTE data. (a, top, left) Cluster 1. Late Up pattern. (b, top, right) Cluster 3. Early Down pattern.(c, bottom, left) Cluster 8. Cycle. Upregulation spike at 8hr.

Discussion

As stated earlier, a quantitative comparison of significant genes in the datasets would not be appropriate given the differences in study design. However, this method of determining significant genes seems to valid, albeit conservative. The genes identified as significant in each dataset are consistent with what is known about both the cognitive decline/aging and brain injury processes. When compared to the original datasets, many of the genes match what was found in the original studies.[5,6]

Qualitatively, the most interesting result pertains to the expression of Cd74, which corresponds to the invariant chain of the major histocompatibility complex II (MHC II). Cd74 plays a diverse role in immune function, serving as both a molecular chaperone for MHC II but also as a receptor in various immune signaling.[12] Cd74 is also known as amyloid beta binding protein. Cd74's role in AD is not well characterized, but primary research has shown that Cd74 inhibits ABO production by directing ABO's precursor, amyloid precursor protein (APP) to endocytic vacuoles.[12] As a result, APP is unable to undergo the proteolytic cleavage necessary to produce ABOs. ABO plaques are what characterize AD disease, although exactly how the plaques contribute to cognitive decline is unknown.

In the AD model, Cd74 was upregulated late in life (**Figure 3b**). In the brain injury model, Cd74 was upregulated at the 8 hour mark, or 8 hours after the injury occurred (**Figure 4c**, see probe name X13044_at). Given the conservative nature of the model, it may be of significance that Cd74 was upregulated in both models, and could indicate some shared pathology between the diseases. It is important to note that neither model is perfect for their representative diseases--the aging model is not unique to AD, and the brain injury model does not model the repetitive damage that is characteristic of CTE. Nevertheless, this is a new finding--Cd74 has never been reported as a potential shared feature of AD and CTE before.[1,3,8]

Conclusion

AD and CTE share a controversial relationship that is not well characterized. Previous research has suggested that they may share pathology related to the genes APOE, MAPT (tau protein) and TDP-43.[1,3,8] However, this is still under much debate. In an effort to explore this debate, we sought to identify significant changes in gene expression over time in both a cognitive decline/aging and traumatic brain injury rat model. Although these are imperfect models, they correspond to both AD and CTE disease states, where a temporal element is important to the pathology of the disease. After identifying significant genes in both datasets, we performed hierarchical clustering on each individual dataset. We discovered some temporal patterns in each dataset that were consistent with what is known about each condition. In particular, we found that Cd74, also known as amyloid beta binding protein, is upregulated in both datasets, but at different time points. In AD, this upregulation is present later in life, around 23 months. In CTE, this upregulation occurs at 8 hours after brain injury, before returning to normal levels after 24 hours. Cd74 has not been identified as a shared feature in these diseases prior to our analysis. However, we could only draw qualitative conclusions on imperfect models. As a result, this finding warrants further investigation and validation. Further research could focus on simultaneously modeling CTE and cognitive decline/aging in a longitudinal fashion. This could show the impact of repeated brain trauma over a longer course of time, in addition to providing comparable datasets for both models. As a result, apples to apples comparisons could be drawn, without worry about confounding effects.

Appendix A:

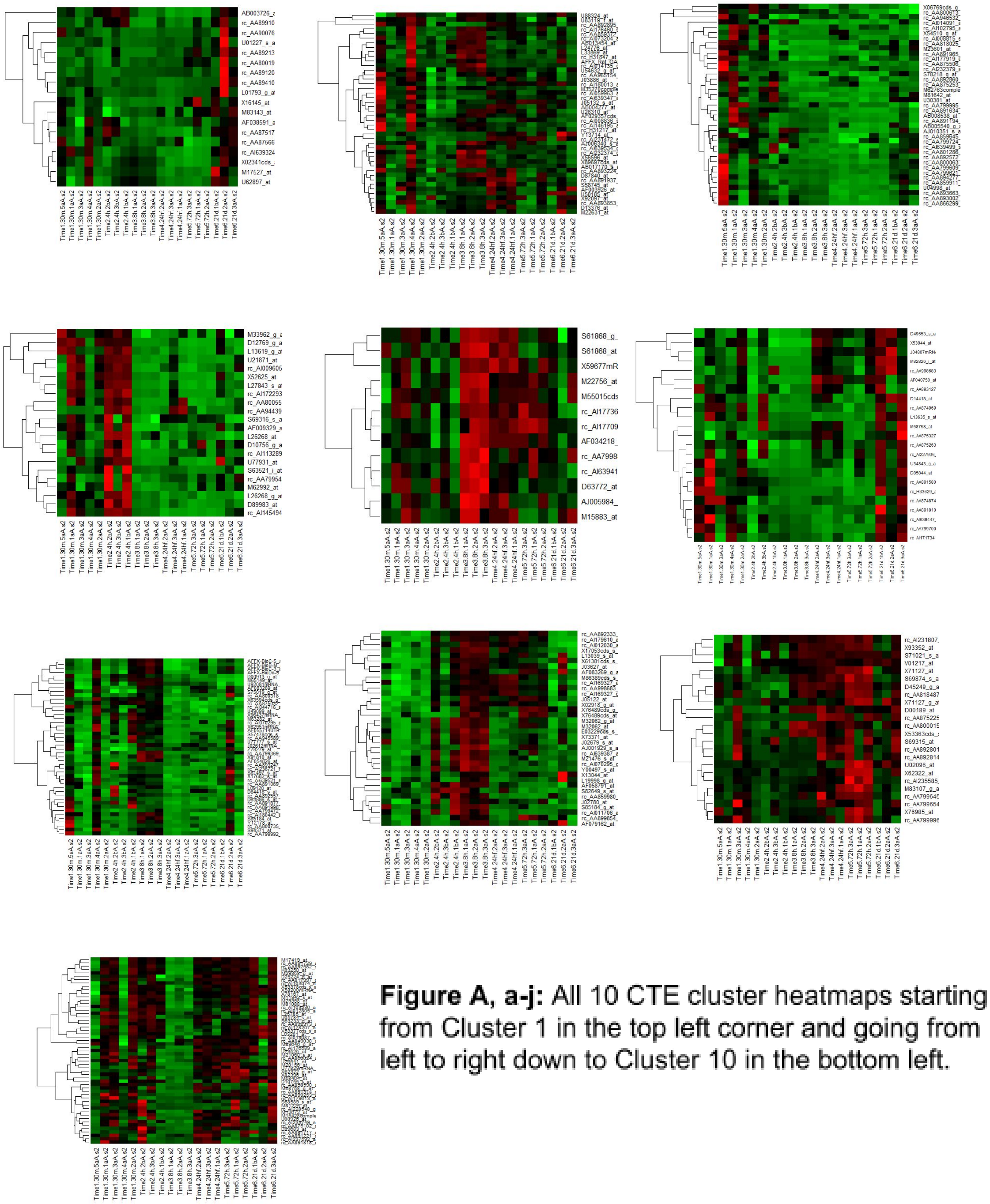

Figure A, a-j: All 10 CTE cluster heatmaps starting from Cluster 1 in the top left corner and going from left to right down to Cluster 10 in the bottom left.

References

1. Armstrong, R. A., McKee, A. C., Stein, T. D., Alvarez, V. E., & Cairns, N. J. (2018). Cortical degeneration in chronic traumatic encephalopathy and Alzheimer's disease neuropathologic change. *Neurological Sciences*, 1–5. https://doi.org/10.1007/s10072-018-3686-6

2. Lehman, E. J., Hein, M. J., Baron, S. L., & Gersic, C. M. (2012). Neurodegenerative causes of death among retired National Football League players. *Neurology*, *79*(19), 1970–4. https://doi.org/10.1212/WNL.0b013e31826daf50

3. Woerman, A. L., Aoyagi, A., Patel, S., Kazmi, S. A., Lobach, I., Grinberg, L. T., … Prusiner, S. B. (2016). Tau prions from Alzheimer's disease and chronic traumatic encephalopathy patients propagate in cultured cells. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(50), E8187–E8196. https://doi.org/10.1073/pnas.1616344113

4. (2016). Chronic traumatic encephalopathy - Symptoms and causes - Mayo Clinic. Retrieved February 13, 2019, from https://www.mayoclinic.org/diseases-conditions/chronic-traumatic-encephalopathy/symptoms-causes/syc-20370921

5. Kadish, I., Thibault, O., Blalock, E. M., Chen, K.-C., Gant, J. C., Porter, N. M., & Landfield, P. W. (2009). Hippocampal and Cognitive Aging across the Lifespan: A Bioenergetic Shift Precedes and Increased Cholesterol Trafficking Parallels Memory Impairment. *Journal of Neuroscience*, *29*(6), 1805–1816. https://doi.org/10.1523/JNEUROSCI.4599-08.2009

6. Natale, J. E., Ahmed, F., Cernak, I., Stoica, B., & Faden, A. I. (2003). Gene Expression Profile Changes Are Commonly Modulated across Models and Species after Traumatic Brain Injury. *Journal of Neurotrauma*, *20*(10), 907–927. https://doi.org/10.1089/089771503770195777

7. A. (2017, March 15). 2017 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *13*(4), 325-373. Retrieved from ScienceDirect.

8. Cline, E. N., Bicca, M. A., Viola, K. L., & Klein, W. L. (2018). The Amyloid- β Oligomer Hypothesis: Beginning of the Third Decade. *Journal of Alzheimers Disease,64*(S1). doi:10.3233/jad-179941

9. Johnson, G. V. (2004). Tau phosphorylation in neuronal cell function and dysfunction. *Journal of Cell Science, 117*(24), 5721-5729. doi:10.1242/jcs.01558

10. 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).

11. Lloyd-Evans, E., & Haslett, L. J. (2016). The lysosomal storage disease continuum with ageing-related neurodegenerative disease. *Ageing Research Reviews*, *32*, 104–121. https://doi.org/10.1016/j.arr.2016.07.005

12. Matsuda, S., Matsuda, Y., & D'Adamio, L. (2009). CD74 interacts with APP and suppresses the production of Abeta. *Molecular Neurodegeneration*, *4*, 41. https://doi.org/10.1186/1750-1326-4-41

13. Liu, W., Mei, R., Di, X., Ryder, T., Hubbell, E., Dee, S., Webster, T., Harrington, C., Ho, M., Baid, J., & Smeekens, S. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12), pp. 1593–1599.

14. Liu, W., Mei, R., Bartell, D., Di, X., Webster, T., & Ryder, T. (2001). Rank-based algorithms for analysis of microarrays. *Proceedings of SPIE, Microarrays: Optical Technologies and Informatics*, 4266.

15. Affymetrix. (2002). Statistical Algorithms Description Document. *Affymetrix Inc., Santa Clara, CA*. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf,

16. Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., & Speed, T. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15

17. Irizarry R., Bolstad, B., Astrand M., & Speed, T. (2003b). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*, 19(2):185-193

18. Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., & Speed, T. (2003c). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, Vol. 4, Number 2: 249-264

19. Y. Hochberg. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*. Vol. 75: 800-802.