Final Project:

Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets

Adrian Koegl, Nicole Pimentel-Soler

Electrical Engineering Department, Columbia University, New York, NY
ECBME 4060: Introduction to Genomic Information Science and Technology
Wei-Yi Cheng, Ph.D.
December 2021

Abstract

Motivation: Through this project we aim to replicate results obtained in the MSBB Study for the subtyping of Alzheimer's Disease (AD) cohorts on a molecular basis as explored through RNAsequencing of tissue samples from brain regions, and the relevance and reliability of these genotypes' correlation to clinical and pathological diagnoses for AD patients.

Results: Subtyping of AD patients in the MSBB cohort proved lack of inter-reliability amongst the factors considered in this approach.

Contacts: aak229@columbia.edu, np2683@cumc.columbia.edu

Supplementary information:
https://github.com/GiantDole/Molecular-Subtyping-of-Alyheimer-s-disease

1 Introduction

Dementia, chronic deterioration of cognitive function beyond what might be expected from biological aging, affects approximately 55 million people worldwide; Alzheimer's Disease (AD) accounts for approximately 70% of these cases (WHO, *Dementia*). Although signs & symptoms associated with AD may be easily recognized through behavior and/or brain imaging, confirmatory diagnosis can only be made through post-mortem pathological analysis (Neff et al., 2021).

The AD brain exhibits extensive mass reduction as a consequence of neurofibrillary tangle and amyloid-beta (Aβ) peptide formations that result in progressed neuronal and synaptic loss (Neff et al., 2021). These distinctive features occur through various pathophysiological mechanisms, yielding patient subsets with distinct clinical frames, and deeming AD a heterogeneous disease (Neff et al., 2021). A series of genetic variants have been linked to a higher predisposition for AD; however, an exact relation between allelic presence and disease expression has not yet been drawn (Neff et al., 2021).

AD heterogeneity, apparent at the molecular level, has prevented scientists from developing therapeutics of reliable effectiveness for the general population living with AD. As we face an aging population, an effective treatment for this leading cause of morbidity and mortality is most imperative. High specificity diagnostic tools and identification of AD associated biomarkers could significantly improve clinical care, prevention, and progress mitigation; all prospects of successes in RNA sequencing.

Through this project we will consider the transcriptome analysis approach undertaken by a research group aiming to identify disease mechanisms and potential targets of AD through RNA sequencing mediated molecular subtyping.

2 Theoretical Background

Although changes in cognition resulting from the aging process are expected and often considered the norm, clinical presentation of AD differs by the wide domain spectrum at which chronic cognitive decline is observed (Hof et al., 2001). Tampered cognitive function has been associated with lesions in the cortical brain region (Hof et al., 2001). However, distinctive AD pathologies may permeate beyond this area, contributing to advanced decline (Hof et al., 2001).

The asymptomatic progression of bio-pathological lesions, and the limitations posed by a post-mortem confirmatory diagnosis, has led to the development of cognitive assessments, psychometric tools, and rating scales for the evaluation of AD. However, specificity of these qualitative resources is not conducive to interpretation of molecular relevance. Therefore, can only offer an approximation of disease progression devoid of representation for AD heterogeneity. Furthermore, definition of functional decline in qualitative diagnostics is contingent on severity at baseline (Atchison et al., 2007; Brown, 2011; Hyman et al., 2012), which may deviate from expected trends due to episodic fluctuations throughout disease progression. Therefore, rate and stage of progression are clinical approximations informed by apparent cognitive domain affect (Hyman et al., 2012) at the time restricted for tool administration. The Clinical Dementia Rating (CDR) is a structured rating scale for the evaluation of memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care cognitive domains independently scored on a 0-5 point scale from which a cumulative score is derived to define AD severity (Hyman et al., 2012).

AD neuropathologies, rich in β-amyloid plaques and Neurofibrillary tangles, precede signs and symptoms at clinical presentation. NFTs result from paired helical filaments of abnormal tau protein (Neff et al., 2021) in the neuronal cytoskeleton (Hof et al., 2001). Early stage NFT's, prevalent in the limbic regions, progressively invade the brain cortex, subcortical nuclei, and brainstem (Hyman et al., 2012; Neff et al., 2021). The Braak staging criterion proposes a six stage series to describe the extent of NFT permeation in the brain: no NFTs, Braak stages I/II with NFTs predominantly in entorhinal cortex and closely related areas, stages III/IV with NFTs more abundant in hippocampus and amygdala while extending slightly into association cortex, and stages V/VI with NFTs widely distributed throughout the neocortex and ultimately involving primary motor and sensory areas (Hyman et al., 2012).

β-amyloid (Aβ) plaques, present in all genetically linked causes of AD, are also definitive bio-pathologies resulting from Aβ- peptide aggregates at the center of dystrophic neurite clusters known as neuritic plaque (Hyman et al., 2012; Vickers et al., 1996). However, differing phenotypes at each brain region increases heterogenic complexities in AD. The Consortium to Establish a Registry for AD (CERAD), a semiquantitative neuritic plaque scoring system, ranking histochemically identified neuritic plaque densities across neocortical regions, classifies patients into one of four categories: AD, possible AD, probable AD, or definite AD, and establishes moderate or frequent neuritic plaque lesions must be present in at least one neocortical region for a neuropathological confirmatory diagnosis (CERAD score; Hyman et al., 2012).

Clinical criteria has recently expanded to encompass patients with milder symptoms, further establishing the need for molecular profiling diagnostics (Yaari et al., 2011). More so with evidence of patients with intact cognition at end of life, displaying substantial AD related neuropathological changes at post-mortem evaluation (Hyman et al., 2012; Yaari et al., 2011). AD's ample spectrum led the National Institute on Aging-Alzheimer's Association (NIA-AA) to support the differentiation between the clinicopathologic term AD, referring to clinical signs and symptoms of cognitive and behavioral changes observed in patients with extensive AD neuropathologic change and AD neuropathologic change, defining three stages of clinical continuum: preclinical, mild cognitive impairment, and dementia- referring to presence and extent of neuropathologic changes observed in autopsy regardless of the clinical setting (Hyman et al., 2012).

Genetic and genome-wide association studies (GWAS) have contributed to the understanding and prevalence of heterogeneity in AD, and although risk factor genes for AD have been identified, these provide limited information on case type, development prediction, and mechanisms of action (Wang et al., 2018).

Considering that established AD risk loci are not indicative of individual risk, investigators for the "Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets" study referred to the Mount Sinai Brain Bank (MSBB) for the production of multi-Omics data in AD and control brains with the purpose of identifying AD molecular basis for application of precision therapeutics (Wang et al., 2018).

The APOE gene has a strong association with AD, and codes for the lipid binding apolipoprotein E protein which maintains healthy cholesterol levels by forming lipoprotein molecules for bloodstream entry (Hyman et al., 2012; Wang et al., 2018). However, the mechanistic function of this protein in relation to AD is still unknown (Serrano-Pozo et al., 2021). The APOE gene presents in three main variants: E2, E3, and E4, where E3 is the most prevalent in the general population (*APOE gene: Medlineplus genetics* 2021) and the E2 variant offers the strongest protective factor (Serrano-Pozo et al., 2021). The role of APOE in AD neuropathologies has expanded from the initial association of the E4 variant to increased β-amyloid plaque formations (*APOE gene: Medlineplus genetics* 2021; Serrano-Pozo et al., 2021). NFTs, microglia and astrocyte responses, and blood brain barrier mechanisms are contributors to the cognitive decline in the AD population and known subjects to this gene's expression (Serrano-Pozo et al., 2021). Despite association with increased risk for AD, presence of the allelic E4 variant is not a direct cause of AD (*APOE gene: Medlineplus genetics* 2021). Similarly, AD occurrence is not exclusive to the E4 variant (*APOE gene: Medlineplus genetics* 2021). Further knowledge of risk variations in the APOE gene will inform mechanistic function in the mitigation and disposition to AD (Serrano-Pozo et al., 2021).

Considering this information, we have decided to recreate a portion of the MSBB study where we consider the RNAsequencing of pathologies observed in the AD cohort for identification of AD subtypes based on gene expression. As in the study of reference, we concentrate our evaluation in data produced for the parahippocampal region given that it contains the highest gene number amongst the regions sequenced for this study and is most likely to provide a comprehensive data set for near accurate evaluation. In line with our established approach, we have chosen a series of variables thought to establish a defined correlation between degree of severity and genotypic expression within these subsets and the prevalence of dominant traits amongst variants- Braak, CERAD, CDR, and APOE genotypes. Through this process we'll develop a machine learning model for predictive classification of AD phenotypes based on the aforementioned considerations. This final product is expected to serve as a potential prototype tool of supreme impact for provision of AD precision medicine, preventative care, and therapeutic development.

3 Methodology

A series of methodological approaches implemented for reproducing elements in the paper of reference, are specified at continuation. Alternate procedures were resourced when addressing challenging and highly complex techniques employed by the authors, while others retained constant with the authors' application.

3.1 Clustering Demented Samples

The goal of our first step was to perform a differential gene expression analysis on three AD subtypes in comparison with nondemented controls. We carried out this analysis on the parahippocampal gyrus (PHG) RNAsequencing data of the MSBB cohort.

First, we loaded the MSBB PHG gene expression data, MSBB metadata, and MSBB biospecimen data into R. The MSBB PHG raw counts were reduced by removing low expressive genes and genes with low variance, leaving us with 17,349 genes from the original 56,632. We then merged the metadata and biospecimen data linking specimen IDs to subject IDs, to map the associated CDR values in the metadata file. All specimens with a CDR == 0.5 were removed as we only considered clinical presentations for evident dementia. Samples with CDR == 0 were defined as the control group. This allowed for the classification of samples in the MSBB gene expression file through the CDR values. Further clustering is described next.

The clustering of samples was obtained by using the cluster_analysis method with hierarchical clustering provided by the multiClust package. WGCNA clustering, similar to the one performed on the paper of reference, was also attempted following the approach in (Clustering using WGCNA); this is provided in a separate R file (see README for more details). We were able to perform a sound gene clustering with this approach and correlate the resulting clusters with apoeGenotype, CERAD, Braak, and CDR, but couldn't obtain reliable sample clustering from this analysis.

3.2 Differential Gene Expression Analysis

To perform Differential Gene Expression Analysis (DEG) we began by fitting all data, including cluster and controls, together. The data was then normalized and converted to a log-scaled (base 2) gene CPM format using voom. The output was fitted to a linear model using lmFit(). We used this linear model for each of the three clusters to fit the contrast between each cluster and control data (CDR == 0) separately. Genes highlighted by these contrasts were ranked using the empirical bayes method, and subsequently, q-values were estimated through the p-values obtained. With the clustering performed, we achieved an output of 3,410 significant genes across all clusters (qvalue < 0.05); similar to what the authors derived in their analysis.

3.3 Machine Learning

Finally, we trained several machine learning models with the labeled MSBB PHG data and carried out testing to identify the best model. In order to train the model we first annotated the reduced raw count data with the clusters we had obtained. Using the DEG output, we further reduced the gene set to encompass significant genes for every cluster. As the number of significant genes per cluster were not equivalent and posed a tendency for unbalanced preference by the machine learning algorithm, we reduced the significance threshold for some of the clusters to obtain approximately 100 significant genes for each cluster. These genes were once again reduced by the mutual_info_classif method, a reduction observed to improve our ML model. From the labeled data, a random 20% set was selected as test data to train several models using Autogluon. Through this process, we obtained 50% accuracy, yet we could achieve an accuracy of about 65% without reducing the genes with respect to their q-values.

4 Discussion

Challenges in reproducibility became quite palpable through this work particularly when attempting to implement the WGCNA algorithm for our clustering. Although we were successful at defining a subtype series, we recurred to a naive clustering method, given the inability of translating the authors' implementation. We believe this limitation had the highest impact on the accuracy extent of our machine learning model, considered subpar by all means definitive of a reliable tool. Another factor which may have contributed to this event could have been the production of inequivalent number of significant genes associated to each cluster, which although balanced through our approach, could have potentially weighted the output. Nevertheless, our work has built a foundation of which output we expect to increase in performance when a more sophisticated clustering algorithm provides the input to our model. In that case, this model can be used to reliably predict the clusters of further gene expression data of the MSBB or ROSMAP cohort as well.

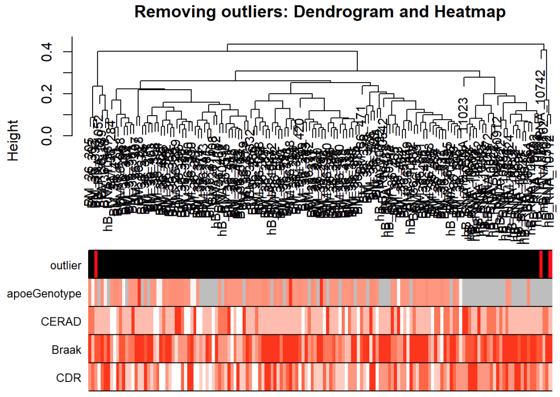

Figure 1: Subject specific trait inter-relation mapping

We also understand that although the APOE genotype is a strong contributor to the definition of AD progression, it is not the only genetic factor in the progression of this disease which is also affected by lifestyle and environmental factors.

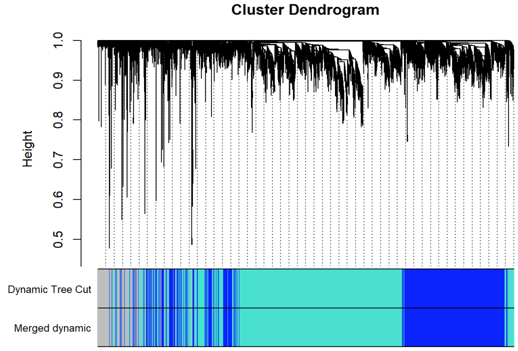

Figure 2: AD cohort cluster dendrogram for subtype identification

The removal of outliers grants us an overview for expression of neuropathologic traits considered in contrast to APOE genotype variance and AD at clinical presentation (Fig.1). As highlighted throughout the theoretical approach for this project, the variability observed in CDR measures is not a sufficient indicator for AD progression considering a marked distinction amongst definitive values for AD as defined by this tool and scales designed for the assessment of bio-pathological lesions does not exist, and appears to have an aleatory relation within this cohort. On the other hand, presence of NFT lesions are seen to counteract β-amyloid (Aβ) plaques, suggesting that there is predominance of pathological expressions in most subjects. Although loosely sustained parameters between lesion types and APOE genotype can be deduced for each subject, the multifunctional association of this gene in the promotion or inhibition of these heterogeneous pathological mechanisms, is not conclusive of and for a particular subtype. This is in part to the multigenic expression in AD. Therefore, we perceive this as a limitation for a comprehensive analysis for interplay between phenotypic expression in this AD cohort and respective RNAsequences considered.

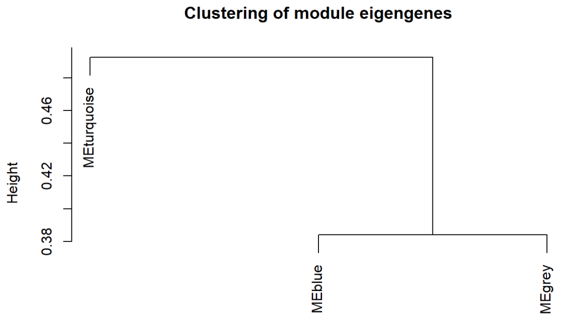

Figure 3: Relative association of AD clusters identified in the MSBB cohort

Despite the limits faced by these considerations, a three subtype series defined by molecular sequencing and trait consideration was attainable (Fig.2), demonstrating an approximation to the results obtained by the authors in the paper of reference. Differing congruence, observed in relative distance between clusters in Fig.3, was demonstrative of module-trait relationships for each cohort (Fig.4), but further proved the lack of interrelation for predictive value within a specific cluster.

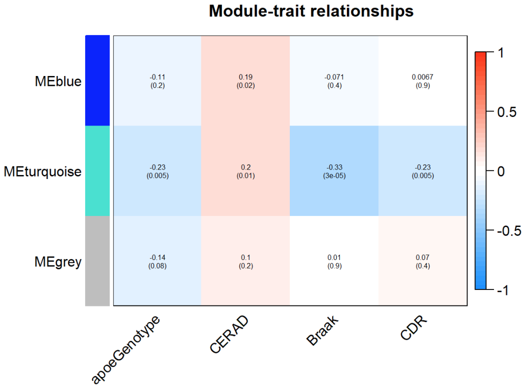

Figure 4: Trait relationships amongst and within identified AD subtypes

We acknowledge the fidelity of these results is not sufficiently well founded, as proven by the accuracy levels obtained in our machine learning model, in part justified by the fact that our analytical approach was grounded on abstraction of published results. Should we have had the knowledge, tools, and time to immerse ourselves in the consideration of alternative sequencing methods and the expanse of applicable analytic approaches, we could have perhaps obtained results surpassing this conjecture.

For the sake of advancement, technique acquisition, and knowledge exchange, we side with the stance and propositions made by Yusuf A. Hannun in regard to facilitating tools for the replication of research results (Hannun, 2021).

Acknowledgements

The authors want to express full gratitude to Prof. Wei-Yi Cheng who has not only presented us with the tools for foundational comprehension in the field of Genomics, but has provided guidance beyond the lecture hall for us to make this work possible. Thank you always for your kindness and benevolent patience.

References

APOE gene: Medlineplus genetics. In: MedlinePlus. https://medlineplus.gov/genetics/gene/apoe/#:~:text=The%20APOE%20gene%20provides%20instructions,carrying%20them%20through%20the%20bloodstream. Accessed 20 Dec 2021

Atchison T, Massman P, Doody R (2007) Baseline cognitive function predicts rate of decline in basic-care abilities of individuals with dementia of the alzheimer's type. Archives of Clinical Neuropsychology 22:99–107. doi: 10.1016/j.acn.2006.11.006

Brown PJ (2011) Functional impairment in elderly patients with mild cognitive impairment and mild alzheimer diseaseimpairment in MCI and AD patients. Archives of General Psychiatry 68:617. doi: 10.1001/archgenpsychiatry.2011.57

CERAD score. In: RADC- Research Resource Sharing Hub. https://www.radc.rush.edu/docs/var/detail.htm?category=Pathology&subcategory=Alzheimer%27s+disease&variable=ceradsc. Accessed 1 Dec 2021

Clustering using WGCNA - bioinformatics team (bioiteam) at the University of Texas. In: UT Austin Wikis. https://wikis.utexas.edu/display/bioiteam/Clustering+using+WGCNA. Accessed 24 Dec 2021

Dementia. In: World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 5 Nov 2021

Hannun YA (2021) Build a registry of results that students can replicate. Nature 600:571–571. doi: 10.1038/d41586-021-03707-9

Hof PR, Mobbs CV (2001) Functional Neurobiology of Aging. Academic Press, San Diego, CA. isbn: 9780123518309

Hyman BT, Phelps CH, Beach TG, et al (2012) National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. Alzheimer's & Dementia 8:1–13. doi: 10.1016/j.jalz.2011.10.007

Neff RA, Wang M, Vatansever S, et al (2021) Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. Science Advances. doi: 10.1126/sciadv.abb5398

Sage Bionetworks info@sagebase.org Sage bionetworks. In: Synapse. https://www.synapse.org/#!Synapse:syn5550382. Accessed 16 Nov 2021

Serrano-Pozo A, Das S, Hyman BT (2021) APOE and Alzheimer's disease: Advances in genetics, pathophysiology, and therapeutic approaches. The Lancet Neurology 20:68–80. doi: 10.1016/s1474-4422(20)30412-9

Vickers JC, Chin D, Edwards A-M, et al (1996) Dystrophic neurite formation associated with age-related β amyloid deposition in the neocortex: Clues to the genesis of Neurofibrillary Pathology. Experimental Neurology 141:1–11. doi: 10.1006/exnr.1996.0133

Wang M, Beckmann ND, Roussos P, et al (2018) The Mount Sinai Cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Scientific Data. doi: 10.1038/sdata.2018.185

Yaari R, Fleisher AS, Tariot PN (2011) Updates to diagnostic guidelines for alzheimer's disease. The Primary Care Companion For CNS Disorders. doi: 10.4088/pcc.11f01262