Transcriptomic profiling of Early-onset sporadic Alzheimer’s disease patients

Andrew B Caldwell
University of California San Diego

Balaji G Anantharaman
University of California San Diego

Srinivasan Ramachandran
University of California San Diego

Phuong Nguyen
University of California San Diego

Qing Liu
University of California San Diego

Ivy Trinh
University of California San Diego

Douglas R Galasko
University of California San Diego

Paula A Desplats
University of California San Diego

Steven L Wagner
University of California San Diego

Shankar Subramaniam (shankar@ucsd.edu)
University of California San Diego  https://orcid.org/0000-0002-8059-4659

Research Article

Keywords:

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1766592/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Version of Record: A version of this preprint was published at Molecular Brain on October 12th, 2022. See the published version at https://doi.org/10.1186/s13041-022-00963-2.
Abstract

Alzheimer’s disease (AD) manifested before 65 years of age is commonly referred to as early-onset AD (EOAD) [1]. Significant research on EOAD has focused on autosomal-dominant familial mutations in PSEN1, PSEN2, and APP; however, these mutations only represent 5–10% of EOAD cases. Although EOAD has a higher heritability (92–100%) than sporadic late-onset AD (LOAD, 70%), which is polygenic, EOAD may be due to autosomal recessive inheritance or combinatorial polygenic risk [2, 3]. Even though the endpoint clinicopathological changes, i.e., Aβ plaques, tau tangles, and cognitive decline, are common across EOAD and LOAD, the disease progression is highly heterogeneous [4]. The heterogeneity, leading to temporally distinct age of onset (AOO) and stages of cognitive decline, may be caused by myriad combinations of distinct disease-associated molecular mechanisms. We and others have used transcriptome profiling in AD patient-derived neuron models of autosomal-dominant EOAD and sporadic LOAD to identify disease endotypes [5–7]. Further, analyses of large cohorts of postmortem brains demonstrate that only one-third of AD patients show hallmark disease endotypes like increased inflammation and decreased synaptic signaling [4]. Areas of the brain less affected by AD pathology at early disease stages—such as the primary visual cortex—exhibit similar transcriptomic dysregulation as those regions traditionally affected, and therefore may offer a view into the molecular mechanisms of AD without the associated inflammatory changes and gliosis induced by pathology [8]. To this end, we analyzed AD patient samples from the primary visual cortex (19 EOAD, 20 LOAD), using transcriptomic signatures to identify patient clusters and disease endotypes. Interestingly, although the clusters showed distinct combinations and severity of endotypes, each patient cluster contained both EOAD and LOAD cases, suggesting that AOO may not directly correlate with the identity and severity of AD endotypes.

Results And Discussion

We sought to identify and characterize molecular phenotypes, or endotypes, associated with Early-onset sporadic (EOS, i.e., cases not caused by autosomal-dominant mutations) Alzheimer’s disease in contrast with Late-onset sporadic (LOS) Alzheimer’s disease using RNA-seq. We selected 40 AD samples from the UC San Diego Shiley-Marcos Alzheimer’s Disease Research Center (ADRC) brain bank stratified into two statistically distinct groups based on the age of onset (AOO) (<60 years, EOAD, 19 samples; >70 years, LOAD, 20 samples after quality filtering) (Fig. 1A). 8 aged, non-demented controls (NDC) were included for comparison. Tissue from the primary visual cortex (Brodmann Area 17, Bm-17) was used for two key reasons: firstly, atypical clinical symptoms, such as visual impairment, are more frequent in EOAD than LOAD; secondly, it mirrors the transcriptomic signature of traditionally affected brain regions without inflammation and gliosis associated with amyloid beta deposition [8,9]. This provides a unique opportunity to understand the AD transcriptomic signature without confounding factors. Following RNA-seq, hierarchical clustering using scaled gene expression signatures and Euclidean distance to compute pairwise similarity scores failed to dichotomize the samples as early-onset (EOS) and late-onset (LOS). Instead, we observed four clusters, only one of which did not have a mixed membership of EOS and LOS cases (Fig. 1B). As the samples did not cluster based on their AOO, we decided to proceed with the 4
clusters based on transcriptional profiles. As postmortem and frozen samples are prone to RNA degradation, we tested the correlation between differential expression (DE) in each cluster with a previously generated reference brain degradation dataset by generating DEqual plots from the quality surrogate variable analysis (qSVA) framework (Fig. 1C) [10]. By this approach, clusters 1, 2 and 4 showed either no or negative correlation between the degradation and AD-induced DE t statistic, whereas cluster 3 showed a positive correlation. Additionally, cluster 3 demonstrated divergence from all other samples in MDS space (Fig. 1D) and had the lowest RNA-seq transcript assignment percentage (Fig. 1E). Therefore, we excluded cluster 3 from further analysis. Interestingly, clusters 1, 2 and 4 did not show a statistically significant difference in AOO or age of death (AOD), although cluster 4 had the earliest AOO and AOD (Fig. 1F). The number of differentially expressed gene transcripts (DEGs) relative to NDC increased from cluster to cluster as the AOO and AOD decreased, with cluster 4 displaying the largest number of DEGs (Fig. 1G). Functional enrichment analysis using the fgsea and CERNO with the GO: Biological Process, Reactome, and Hallmark databases (Fig. 1H) revealed gene sets related to dedifferentiation and non-ectoderm lineage definition, inflammation, synaptic function, and oxidative phosphorylation. Enrichment using the StringDB v10 database demonstrated activation of genes that are protein-protein interaction (PPI) partners with TGFB signaling (TGFB1, CTNNB1), transcription factors (TFs) which activate EMT/dedifferentiation (YAP1, WWTR1/TAZ), as well as proteins previously implicated in AD (SRC, SEC61G, EEF2, RPL7) [11–13]. Next, we used ISMARA motif activity analysis to find TFs with differential activity across AD clusters. This revealed activation of TFs controlling early-stage neural lineage commitment or repression of neuron specification and function (REST), repression of neuronal mitochondrial energy production (NRF1) and other neural factors (MEIS2, ZNF711) (Fig. 1I). Further, TFs involved in non-ectoderm/precursor lineage (TEAD1, SPI1/Pu.1, SNAI2), inflammation (REL, IRF1/8), chromatin modulation (EZH2, MTA3, MECP2), and pluripotency (KLF4, GATA3) were also enriched, particularly in cluster 4. Next, we wanted to identify co-expressed gene modules differentially regulated across the AD clusters. We performed CEMiTool analysis using the CEMiTool R package for all genes commonly expressed in AD and NDC samples with >10cpm normalized expression [14]. Twenty-two co-expression modules and one non-correlated module were identified from the 9120 genes, ranging from 48 to 2062 genes in size. Enrichment scores for each co-expressed module were calculated for each AD and NDC sample using Gene Set Variation Analysis (GSVA) in the GSVA R package (Fig. 1J) [15]. The ontological identity of each module was characterized by hypergeometric enrichment of module genes with GO: BP, Reactome, and Hallmark databases (Fig. S1) and their statistical significance by enrichment with camera [13] method. This revealed activation of modules functionally associated with non-ectoderm dedifferentiation and early neurogenesis (M1, M3) and chromatin modification (M2, M8), as well as repression of modules associated with neuron lineage and function (M1, M4, M6, M9), and oxidative phosphorylation (M4, M10, M18), particularly in clusters 2 and 4 (Fig. 1K). Using the similarity dendrogram of module GSVA scores across all samples, we merged individual modules into comodules based on GSVA score similarity and common ontological categories. Three comodules in particular, M6-M9, M5-M11-M14, and M4-M10-M18, had closely related cellular functions and stratified the three AD clusters (Fig. 1L-M). Comodule M6-M9 was significantly enriched for synaptic signaling and neuron differentiation with the neural repressor REST and non-ectoderm lineage factor SMAD4 as key regulators.
M6-M9 genes were substantially downregulated in cluster 4, modestly downregulated in cluster 2, and mixed regulation in cluster 1. In contrast, comodule M5-M11-M14 – which contains genes involved in cell cycle and proliferation (MYC targets), membrane trafficking, and oxidative phosphorylation – was upregulated in clusters 1 and 4 but downregulated in cluster 2. Interestingly, comodule M4-M10-M18, whose top transcriptional regulators are NRF1 and CREB1, is also enriched for oxidative phosphorylation and synaptic signaling as observed in the two other comodules. However, M4-M10-M18 genes were most downregulated in cluster 2, suggesting that while loss of synaptic function and oxidative phosphorylation are common between clusters 2 and 4, distinct pools of genes are differentially regulated in the two clusters. Further, their expression loss is mediated by unique regulators (e.g., gain of repression by REST versus loss of activation by NRF1).

At the outset, we aimed to characterize the relationship between the severity of AD endotype dysregulation with the Age of Onset (AOO). However, we quickly determined that the patient transcriptional profiles in our AD cohort did not bifurcate into early-onset (EOS) and late-onset (LOS) cases. Following filtering, three clusters were identified with differing ratios of EOS and LOS patients and average AOO. Cluster 2 and 4 were closer in terms of average AOO and enrichment of transcriptional programs, or endotypes, differentially modulated relative to NDC; ontological enrichment and TF activity suggest that cluster 4 may represent a more severe form of cluster 2. In contrast, cluster 1 demonstrated a later average AOO concomitant with the severity of co-expression modules or disease endotypes that were either lesser than the other two clusters (e.g., lower reactivation of REST or the M5-M11-M14 comodule) or in the opposite direction of other clusters (e.g., M4-M10-M18 comodule and synaptic function). The observation that some patient clusters within AD cohorts exhibit transcriptional dysregulation that are opposite of hallmark AD changes (e.g. loss of synaptic signaling, activation of inflammation) has been described before in larger studies [4]. Interestingly, cluster 1 only contained one patient with an early-onset diagnosis; it is possible that with a larger cohort, the AD profile captured by cluster 1 would have a more substantial contribution of LOS cases that would statistically separate from clusters with a strong representation of EOS cases (e.g., the cluster 4 AD profile). We anticipated that the primary visual cortex, a brain region less affected by AD pathology, may exhibit fewer changes; surprisingly, they reflect analogous dysregulation of disease endotypes observed in other regions albeit with limited contribution from inflammation upregulation. In our previous study of iPSC-derived neurons from patients harboring EOFAD-causing \textit{PSEN1} mutations, we observed a stronger enrichment and simultaneous combination of common endotypes in mutations associated with earlier AOO (i.e., \textit{PSEN1}^{M146L}) [5]. We surmise that the same trend may hold for early-onset as well as late-onset sporadic AD. However, further characterization of specific cell types (e.g., single cell RNA-seq), additional profiling of gene and regulatory mechanisms (e.g., ATAC-seq, genome sequencing), and larger cohorts will be necessary to disentangle the complexity of gene regulatory control of disease endotypes arising from heterogenous EOAD and LOAD mechanisms.

\textbf{Declarations}
Ethics approval and consent to participate

Brain samples were obtained at the UC San Diego Shiley-Marcos Alzheimer’s Disease Research Center (UCSD ADRC) in accordance with UC San Diego IRB approval.

Availability of data and materials

RNA-seq data is available at the NCBI GEO under the accession GSE203206. RNA-seq counts and pathology metadata can be found in Supplementary files 1 and 2.

Competing Interests

The authors declare no competing interests. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Funding

This study was supported by NIH grants P50 AG05131 (D.R.G.), U01 NS 074501-05 (S.L.W.), R01 LM012595 (S.S.), U01 CA198941 (S.S.), U01 DK097430 (S.S.), R01 HD084633 (S.S.), and R01 HL106579-07 (S.S.); NSF grant STC CCF-0939370 (S.S.); Veterans Affairs RR&D 1I01RX002259 (S.L.W); and Cure Alzheimer’s Fund (CAF) grants to S.L.W.

Author Contributions

A.B.C. and B.G.A. carried out the data processing and analysis of the RNA-seq data. A.B.C. wrote the manuscript and generated the figures with contributions from B.G.A. and S.R.. I.T. performed brain tissue sample dissection. P.A.D., P.N., and Q.L. carried out the RNA extraction from brain tissue. A.B.C., D.R.G., S.L.W., P.A.D., and S.S. were involved in the study design and patient selection; S.S. supervised the overall study design, analysis, and manuscript revisions.

Acknowledgments

RNA-sequencing was conducted at the IGM Genomics Center, University of California, San Diego, La Jolla, CA (P30DK063491, P30CA023100, P30DK120515).

References


Methods

Postmortem brain samples

Alzheimer’s disease (AD) and healthy, non-demented control (NDC) samples were obtained from the brain bank of samples preserved at the UC San Diego Shiley-Marcos Alzheimer’s Disease Research Center (UCSD ADRC), extracting tissue from the Brodmann Area 17 (Bm-17) of the occipital lobe (OL) in accordance with UC San Diego IRB approval. A total of 50 samples were selected: 8 NDC and 40 AD patient samples. The 40 AD samples were selected based on their lack of alternative diagnosis (e.g. Lewy Body Dementia, hippocampal sclerosis), APOE status (all AD and NDC samples were either APOEε3/3 or the APOEε3/4 genotype), and stratified into two groups based on the age of onset of AD: early-onset, i.e. those with an age of onset less than 60 years (n=19), and late-onset, i.e. those with an age of onset between 70 and 80 years (n=21). Three cognitive evaluation scores, BIMC (Blessed Memory Information Concentration) [16], MMSE (Mini-Mental State Examination) [17], and Mattis’ DRS (Dementia Rating Scale) [18] were used to classify the selected patients as AD or NDC condition. All NDC patients had a BIMC score <= 4, MMSE score between 26 and 30, and an aggregate DRS score between 127 and 140. Each brain sample was also staged based on the concentration of Neurofibrillary Tangles (NFTs) in different brain regions, using a modified version of the staging scheme introduced by Braak and Braak. All AD samples were classified as BRAAK stage VI, while the NDC samples were classified at BRAAK stage I or II. Additional metadata for each sample was also collected for each AD sample: sex, age of onset, age at death, and the concentration of neuritic plaques and tangles in the mid-frontal cortex (MF), inferior parietal cortex (IP), superior temporal cortex (ST), and hippocampus. As AD was ascertained to be the cause of death of all patients within this study, disease-specific survival (DSS) time was estimated by subtracting the age at diagnosis from age at death.

RNA sequencing
RNA from brain samples was extracted using the RNeasy Lipid Tissue Mini kit (Qiagen Cat. 74804) according to the manufacturer’s protocol. Libraries were prepared for RNA-Seq using the TruSeq Stranded Total RNA Library prep kit (Illumina, Cat. RS-122-2303) by the Ribo-Zero ribosomal RNA reduction method (Illumina). Samples were sequenced at the UC San Diego Institute for Genomics Center sequencing core on an Illumina HiSeq4000 generating Paired-End, 75bp reads with an average of 100 million reads per sample (Illumina, Cat. FC-410-1001).

**RNA-seq data processing and sample clustering**

RNA-Seq data preprocessing was performed using the TrimGalore! package [19], removing sequencing adaptors and selecting for all paired-end reads above a quality score threshold (Phred Q>20). Trimmed RNA-Seq reads were mapped to the GRCh38.p12 human transcriptome using kallisto v0.46.1 [20] with the options -bias and --rf-stranded. The R package tximport v1.8.0 [21] was used to summarize kallisto transcript abundancies to the gene level. A DGEList object was created from gene-level read counts using the DGEList() function from edgeR v3.30.3 [22]. Gene-level count filtering was applied using the filterByExpr function in edgeR for inclusion in further analysis, followed by count normalization using the TMM (Trimmed Mean of M-values) method using the function calcNormFactors. Hierarchical clustering was applied using the factoextra R package [23] to identify clusters of AD patients with similar transcriptional profiles. Genes were filtered by applying 10 counts per million (cpm) minimum threshold across all samples and expression corrected for sex using the removebatcheffects function in the limma R package. and Euclidean distance to compute pairwise similarity between the samples used to compare the dendrograms that ensure from either clustering analysis. The voom function from the limma v3.44.1 R package [24,25] was used to model the mean-variance trend and capture gene-specific weights, which were subsequently used to fit a linear model to the count data including sex and RIN score. A contrast matrix was used to compare gene expression between the AD cluster subtypes and NDC samples, and empirical Bayesian statistics for the differential expression analysis was estimated using the eBayes function from limma. Genes with an FDR-adjusted p-value of less than 0.05 were deemed as being differentially expressed between each AD condition and NDC. To determine whether quality surrogate variable analysis (qSVA) would be useful to apply to the AD cluster subgroups to correct for degradation associated with RIN score, DEqual plots, a diagnostic plot that shows the correlation in differential expression t-statistics between AD-induced and degradation-induced differential expression, were generated to quantify degradation in different sample clusters [10].

**Enrichment and transcription factor activity analysis**

Geneset enrichment analysis was performed by two weighted approaches: competitive, directional enrichment using the fgseamultilevel function in the fgsea [26] R package and non-competitive, unidirectional enrichment using the tmocERNOTest function in the tmoc [27] R package, both with The Gene Ontology-Biological Process (GOBP) [28], Reactome [29], and Hallmark Pathway [30] ontology geneset databases as well as the StringDB [31] protein-protein interaction (PPI) database. For fgsea, genes were ranked by the limma t-statistic, while for CERNO genes were ranked by minimum significant
distance (msd). Transcript reads for ISMARA motif activity analysis were filtered using the `filterByExpr` function in `edgeR`. To determine a directional z-score for each enriched motif identified, the differential z-score for each given motif between each AD cluster and NDC was multiplied by the sign of the Pearson correlation between each motif and its target genes. In cases where ISMARA did not calculate a Pearson correlation, literature evidence of the activator or repressor function of the given TF was used.

**Co-expression module analysis**

Modules of co-expressed genes across NDC and AD samples (clusters 1, 2, and 4) were identified using the `CEMiTool` R package [14] on genes with >10cpm expression across all samples. Prior to module detection, counts were transformed using the `voomwithqualityweights` function in the `limma` R package with the parameters directed=TRUE and cor_method=pearson. Gene Set Variation Analysis (GSVA) [15] using the `GSVA` R package was subsequently performed on the resulting 22 co-expression modules across NDC and AD samples. To determine the enrichment of CEMiTool modules in each AD cluster relative to NDC, GSVA scores were used as an input for the `camera` [32] enrichment function in the `limma` R package. Hypergeometric enrichment of CEMiTool modules or comodules was performed using the `tmodhgtest` function in the `tmod` R package [27] and Gene Ontology: BP [28], Hallmark [30], or Reactome [29] ontology gene set databases as well as ENCODE-ChEA Consensus [33,34] and ReMap [35] TF-gene target databases. The StringDB PPI database v10 [31] was filtered for high-confidence interactions sourced from a) databases and b) literature physical interactions and subset for the genes in a given module.

**Figures**
A. Sex, age of onset, and age of death of early-onset sporadic AD (onset age < 60y), late-onset sporadic AD (onset age > 70y), and non-demented control (NDC) patient Occipital Lobe samples. 

B. Cluster dendrogram of all AD samples based on the expression of genes (8934) with 10cpm across all samples. 

C. DEqaul plot of correlation between differential expression (relative to NDC) and reference patient brain RNA degradation in the 4 clusters. 

D. Multi-Dimensional Scaling (MDS) plot of all patient samples for the
top two dimensions. E. Transcriptome assignment % (kallisto) across the sample groups. F. Age of onset (left) and age of death (right) across the sample groups. G. RNA-seq Volcano plots for the three AD clusters. H. Ranked enrichment analysis of gene expression signatures for the three AD clusters using the GOBP, Hallmark, Reactome, and StringDB databases by the *tm*od CERNO (left) and *fgsea* enrichment test (right). I. TFs with predicted significant activity change by ISMARA motif analysis curated into canonical ontological categories. J. GSVA heatmap, dendrogram, and gene size of the 22 co-expression modules identified by CEMiTool across the sample groups. K. Camera enrichment analysis of the 22 co-expression modules in the three clusters relative to NDC. L. StringDB PPI interaction networks for ontologically- and expression-related comodules across the three clusters; genes color-coded by limma t-value. M. Top enriched pathways and TFs for each comodule (hypergeometric test).

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [220615SubramaniamADRCSupplementaryFigure.pdf](#)
- [220615Subramaniam.ADRCBrainMetadata.tsv](#)
- [220615Subramaniam.ADRCBrainRNAcounts.tsv](#)