Transcriptome Meta-Analysis Suggests the Existence of Two Molecular Subtypes in Alzheimer's Disease

Anahita Nazari
University of Tehran
Sayed-Amir Marashi (marashi@ut.ac.ir)
University of Tehran, College of Science https://orcid.org/0000-0001-9801-7449
Hesam Montazeri
University of Tehran

Research

Keywords: Transcriptome heterogeneity, Subtyping of Alzheimer's disease, Molecular signature

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-466388/v1

Transcriptome meta-analysis suggests the existence of two molecular subtypes in Alzheimer's disease

Anahita Nazari [1], Sayed-Amir Marashi [1]*, Hesam Montazeri [2]

[1] Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran
[2] Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran

*Corresponding author
E-mail address: marashi@ut.ac.ir (S.-A. Marashi)

Abstract:

Alzheimer's disease (AD) is the most common type of dementia that affects the lives of nearly 50 million people globally. No efficient treatment or cure has been found for AD yet. Different studies have been performed on heterogeneity in AD from various perspectives. Analyzing AD as a disease with several subtypes might assist us to identify effective personalized treatments for AD patients. Here, we investigated the heterogeneity in AD from a molecular perspective. In particular, we used gene expression profiles of 93 samples from the hippocampus of AD patients. To increase the sample size and enhance the robustness of the results, we applied meta-analysis on three different datasets. We first combined gene expression profiles from multiple relevant AD studies, and then, using a clustering analysis, we found two distinct molecular subtypes across AD samples. We validated the obtained results using various statistical and randomization approaches. Furthermore, we showed that the two transcriptomic subtypes are statistically independent of sex, age and Braak stage of the subjects. Notably, one of the identified subtypes indicates high similarity to control samples, which suggests that AD onset does not necessarily affect the hippocampal transcriptome. By performing differential gene expression and pathway enrichment analyses for various settings, we found 2371 differentially expressed genes that can be used for discriminating between the two AD subtypes. Finally, using a clustering approach, we introduced a novel molecular signature for the two subtypes based on14 differentially expressed genes.

Keywords: Transcriptome heterogeneity; Subtyping of Alzheimer's disease; Molecular signature

Introduction

Alzheimer's disease (AD), as a progressive neurodegenerative disease, is the most common type of dementia [1, 2]. Memory loss, attention problems, deficits in language and visuospatial abilities are some of the symptoms that patients experience [3–5]. Despite its prevalence, there is currently no efficient treatment or cure for AD [6]. Several different studies have shown that AD is in fact clinically, pathologically, and biochemically heterogeneous [7, 8]. The present "incurability" of AD is suggested to be linked to its heterogeneity [9]. A prominent possibility is that AD consists of multiple subtypes with different mechanisms that need to be treated using different approaches.

There are different viewpoints for subtyping AD. Based on the classical "staging hypothesis", the spread of neurofibrillary tangles initiates from the entorhinal cortex and progressively occupies the association cortex. According to this hypothesis, there are different stages of AD severity, namely the Braak stages I to VI [10, 11]. In some recent studies, the staging hypothesis is challenged and a new "distinct subtypes hypothesis" is recommended. In this viewpoint, AD samples are categorized, based on brain atrophy patterns, into different subtypes [12–15] . The age of onset [16] and cognitive symptoms [17, 18] are some other factors, based on which subtyping can be performed. However, there is currently no universal consensus on subtyping AD.

The complex nature of AD suggests the existence of multiple distinct subtypes and mechanisms at its molecular roots. To gain a better understanding of this disease, there have been several AD studies analyzing molecular data such as transcriptome [19–21]. Some studies focused on a specific subset of genes (such as synaptic genes in [22] and diabetes-related genes in [23]), while others focused on expression levels of all genes [24]. Molecular subtyping has been previously used for studying complex diseases such as cancer [25, 26].

Analyzing post-mortem gene expression profiles obtained from brain regions can provide us with precious information for studying molecular variation during AD [27]. Of particular, the hippocampus has been investigated in several transcriptome studies of AD [28–30] since it is one of the early brain regions affected in this disease [31, 32]. Furthermore, changes in the hippocampus (such as hippocampus atrophy) are associated with AD and some of its symptoms such as memory loss [33, 34]. There was one study on molecular subtyping using RNA-seq data from different brain regions, that showed the hippocampal area demonstrates the greatest subtyping signal over the other regions [35].

In the present work, we used transcriptome data of hippocampus to investigate the existence of multiple AD subtypes. Limited sample size, poor reproducibility and low robustness are well-known problems in the transcriptome analysis of AD. Integrating multiple datasets and performing meta-analysis increases the sample size, which in turn, may lead to more reliable conclusions [36–38]. Meta-analysis of AD expression profiles was previously used for differential

analysis of AD and healthy subjects or for comparison of different severities of AD [39, 40]. Here, we incorporate this strategy for identifying the two different subtypes of AD. To this end, gene expression profiles of the hippocampus from three AD studies were combined and analyzed to gain a better understanding of molecular heterogeneity in AD.

Materials and Methods

***Datasets*:** Hippocampus is a crucial component of the medial temporal lobe memory system. It is generally believed that at the onset of AD, neuronal loss and synaptic and intraneuronal molecular remodeling occur in the hippocampus [32]. Therefore, we focused on gene expression (GE) profiles of post-mortem hippocampus from AD subjects. From the GEO database [41], we found three relevant microarray datasets. We restricted our analysis to hippocampus samples of the datasets, which includes 93 AD and 25 control samples (Table 1).

Table 1: A summary of gene expression data used in the present work.

	GEO accession number	Platform	No. of AD samples used in the present work	No. of control samples used in the present work	Reference
Dataset 1	GSE84422	Affymetrix	55	0	[42]
Dataset 2*	GSE29378	Illumina	16	16	[43]
Dataset 3	GSE1297	Affymetrix	22	9	[44]
Total			93	25	

** In this dataset, many samples were taken in pairs from two different regions in the hippocampus (namely, CA1 and CA3). In such cases, the averaged GE profile was considered for each subject.*

Code availability: All data analyses were performed in the R environment. The codes are available in *Additional file 1*.

***Combining datasets*:** Dataset 1 and 3 are Affymetrix microarray datasets, while Dataset 2 is an Illumina microarray dataset. In order to integrate these three datasets, we performed the preprocessing steps shown in Figure 1. Part of the pipeline is adapted from previously published methods [17], [26].

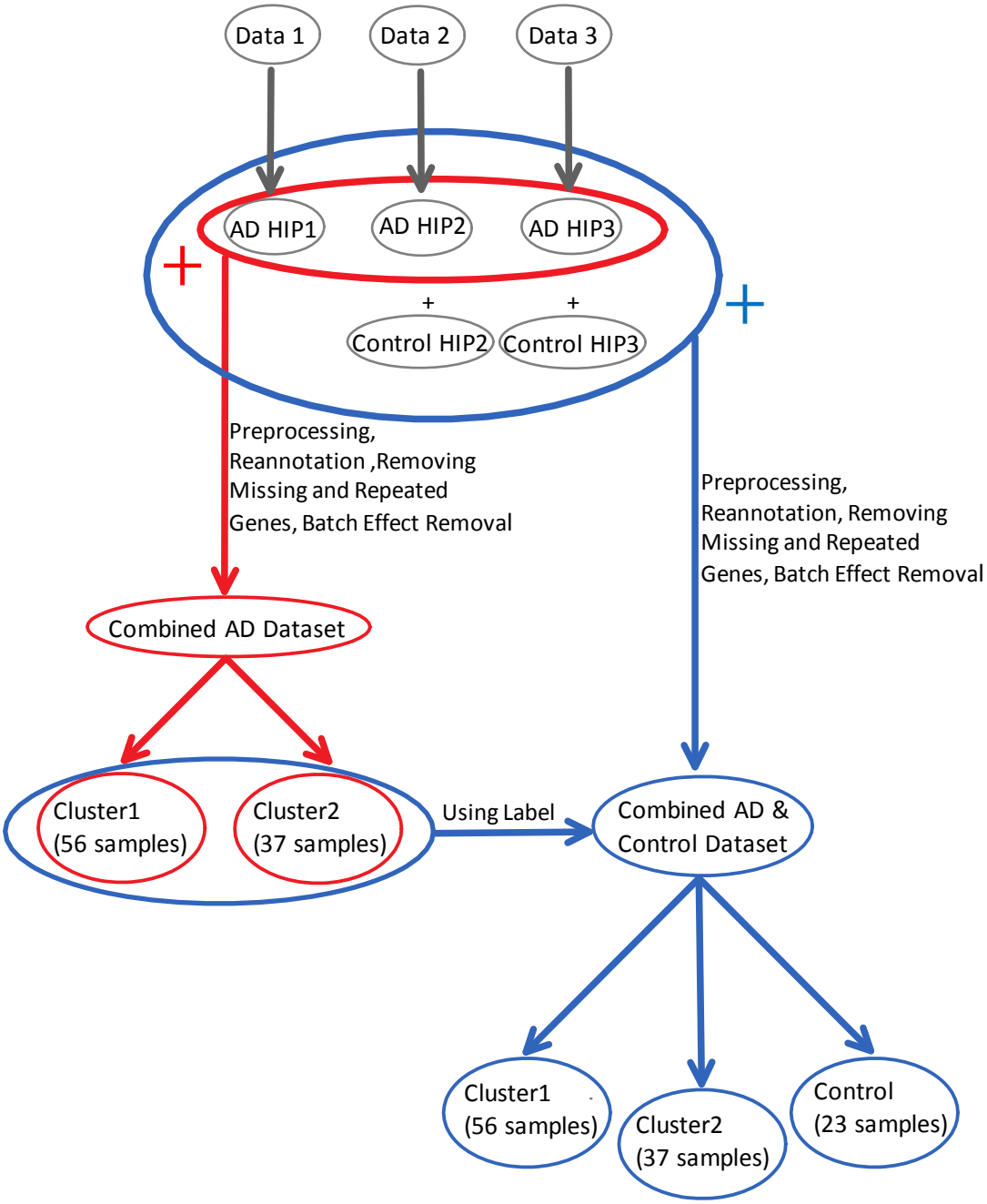

Figure 1: Schematic representation of the data processing for normalizing and combining the three datasets. The combining protocol is based on previously published methods [38, 45]. Briefly, Affymetrix data were preprocessed by 'rma' function in 'affy' package [46] and re-annotated by biomaRt [47]. The Illumina data were "$\log_2$ transformed", and then, normalized (by quantile normalization approach from package 'preprocessCore' [48]) and re-annotated by biomaRt. After removing gene redundancies, for the combined dataset, batch effect removal was performed by ComBat from 'sva' package [49]. (Red steps: combining only AD samples to find AD clusters; Blue steps: combining AD and controls samples and applying the cluster labels obtained in the "red steps").

Determining the optimal number of clusters:

To the best of our knowledge, Dataset 1 [42] is currently the largest available dataset of microarray GE profiles of hippocampal samples from AD subjects. We investigated whether AD samples in this dataset can be clustered into statistically significant subgroups, which potentially represent the existence of multiple AD subtypes. To achieve this goal, we assumed that the data comes from $k = 2, 3, \ldots, 10$ subtypes, and hence, the dataset should be partitioned into k clusters. We used function 'pam' in package 'cluster' [50] to compute average silhouette value, as a measure of clustering validity [51]. The value k that maximizes the average silhouette value of the partitioning was chosen as the optimal number of clusters. We also used function 'pam' in package 'cluster' to determine the clusters [50] . In our case, $k=2$ was found to be the optimal value for clustering the data.

Statistical Analyses:

Details about statistical tools and analyses are presented in Additional file 1. In the following, we present an overview of statistical tools and analyses used in the present work.

Finding DEGs in the combined dataset

The 93 AD samples of our combined dataset were clustered into two groups: cluster 1 which comprises 56 samples, and cluster 2 which comprises 37 samples. Then, we combined our three datasets again, but this time, instead of only using AD samples, we also included healthy samples in all of our steps (including batch effect removal). Next, we partitioned our samples into three groups, one healthy group and two AD groups consisting the clusters that we found in the previous step. We determined the DEGs between the clusters and healthy samples using 'limma' package [52]. A DEG was considered statistically significant if its FDR was smaller than 0.05 and its absolute log fold change larger than 0.4.

Validation

To assess the reproducibility of the results obtained by our method, we repeated the above steps on the largest dataset (Dataset 1) to determine those genes that are expressed differentially between the two clusters of Dataset 1.
Then we investigated whether the identified DEGs and overlaps are statistically significant. In particular, we compared the DEG analysis results to those of random permutations. To this end, two different randomization procedures were considered:

1. Randomizing Dataset 1 via shuffling the expression profile of each gene across samples. This step is repeated 10,000 times.
2. Randomly partitioning Dataset 1 with subgroup sizes equal to the sizes of the genuine clusters. This step is repeated 10,000 times.

To check if the statistically significant subgroups from Dataset 1 are observed in other datasets as well, firstly Datasets 2 and 3 were combined (see Figure 1). This combined dataset will be referred to as Dataset 2+3. Then, clustering of Dataset 2+3 was performed based on the GE profiles of those genes which were significantly upregulated in Dataset 1. Next, we compared the sets of DEGs that were found in Dataset 1 and Dataset 2+3. A good agreement between the two sets is expected to be observed if Dataset 2+3 includes a similar set of subgroups with comparable GE profiles.

To confirm that the agreement between the sets of DEGs is not by chance, Dataset 2+3 was randomly partitioned, with subgroup sizes equal to the sizes of the genuine clusters. Then, we compared the number of significant DEGs between the pair of clusters with those of the genuine clusters. This step is repeated 10,000 times.

Gene set enrichment analysis:

In order to understand the difference between AD subtypes at the molecular level, we analyzed the set of DEGs. To this end, we proceeded by the following steps. Firstly, all AD and control samples in Datasets 1, 2 and 3 were combined, as explained above (Figure 1). Afterward, we identified DEGs between each AD subtype and control as well as between the two AD subtypes. Finally, we used DAVID [53] to perform gene set enrichment analysis for obtained DEGs in each case.

Identification of molecular signatures:

We determined a “molecular signature” that can be used as a set of markers for distinguishing the two AD subtypes. To achieve this goal, we used the identified gene clusters by the biclustering analysis. Each gene cluster represents a set of genes with comparable GE profiles. Then, in each cluster, the medoid gene (*i.e.*, the gene whose average dissimilarity to all the genes in the same cluster is minimal) was chosen as the representative of that cluster to be in the molecular signature.

Results and Discussion:

Clustering of the combined dataset and obtaining DEGs

The average silhouette width suggests that the best partitioning occurs at k=2 clusters (Figure 2). Based on this result, we suggested a “two-subtype model” for hippocampal GE profiles in AD patients.

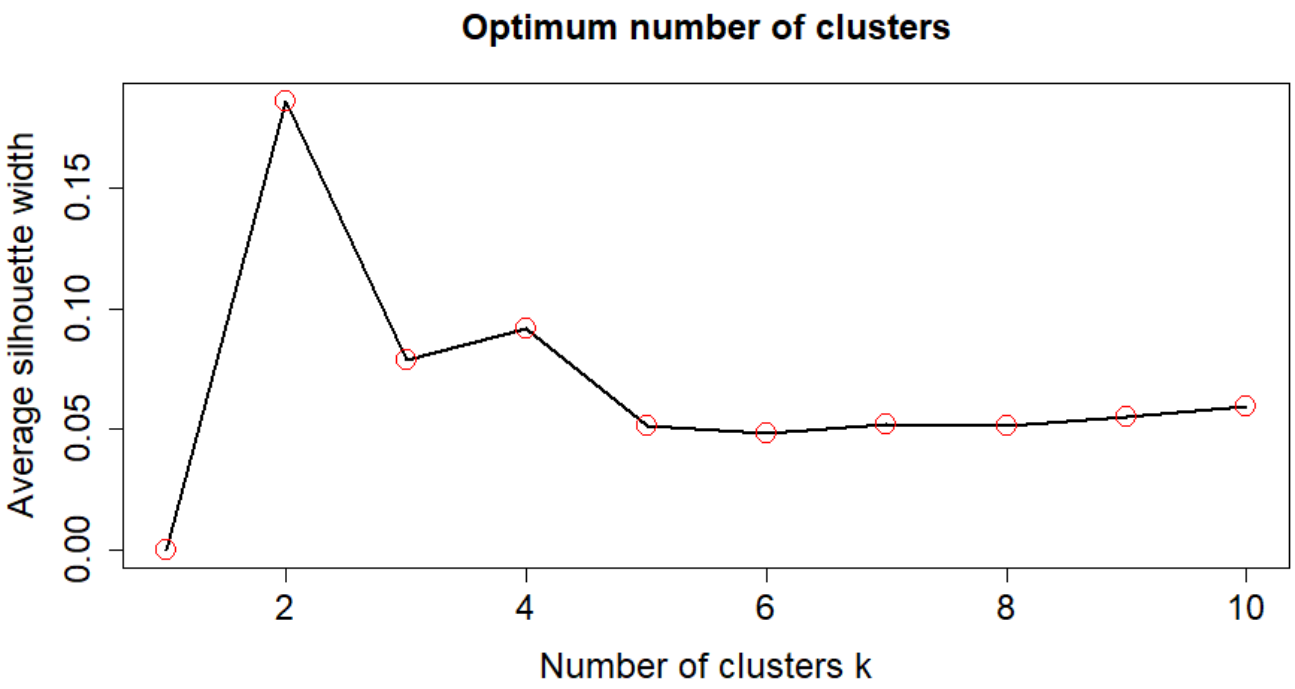

Figure 2: Average silhouette width for partitioning Dataset 1 into *k*=2,3,…,10 clusters

We demonstrated that the GE profiles of the three AD datasets show two distinct patterns, which may represent two different subtypes of AD. In the next step, we combined the AD samples of the three datasets as explained in the Materials and Methods section. The combined dataset will be referred to as “Dataset A”, in which genes with the largest variations across samples were determined. Figure 3a shows the top 300 genes with the greatest variation across samples. After batch effect removal, different datasets are practically indistinguishable, *i.e.*, samples of different datasets are mixed. Based on 14427 mutual genes in the three datasets (after reannotation to Ensembl gene ID), Dataset A was portioned into two subgroups, that is, cluster “c1” with 56 AD samples and cluster “c2” with 37 AD samples (Figure 3b). Interestingly, the two clusters showed no significant association with gender, age and Braak staging (see Additional file 2). Altogether, when only analyzing AD samples 1558 genes had a significantly higher GE level in c1 compared to c2, while only 88 genes had significantly lower GE levels in c1 compared to c2 (FDR<0.05, LFC>|0.4|). To check the significance of the identified DEGs, we randomly partitioned the 93 samples of Dataset A into two subgroups with 56 and 37 samples. This procedure was repeated 10,000 times and the number of genes with significantly higher (respectively lower) GE levels in c1 compared to c2 was counted. Notably, in none of these permutations, the number of genes with significantly higher GE levels in c1 exceeded 1558. Furthermore, only in 12 out of 10,000 permutations, the number of genes with significantly lower GE levels in c1 was larger than 88.

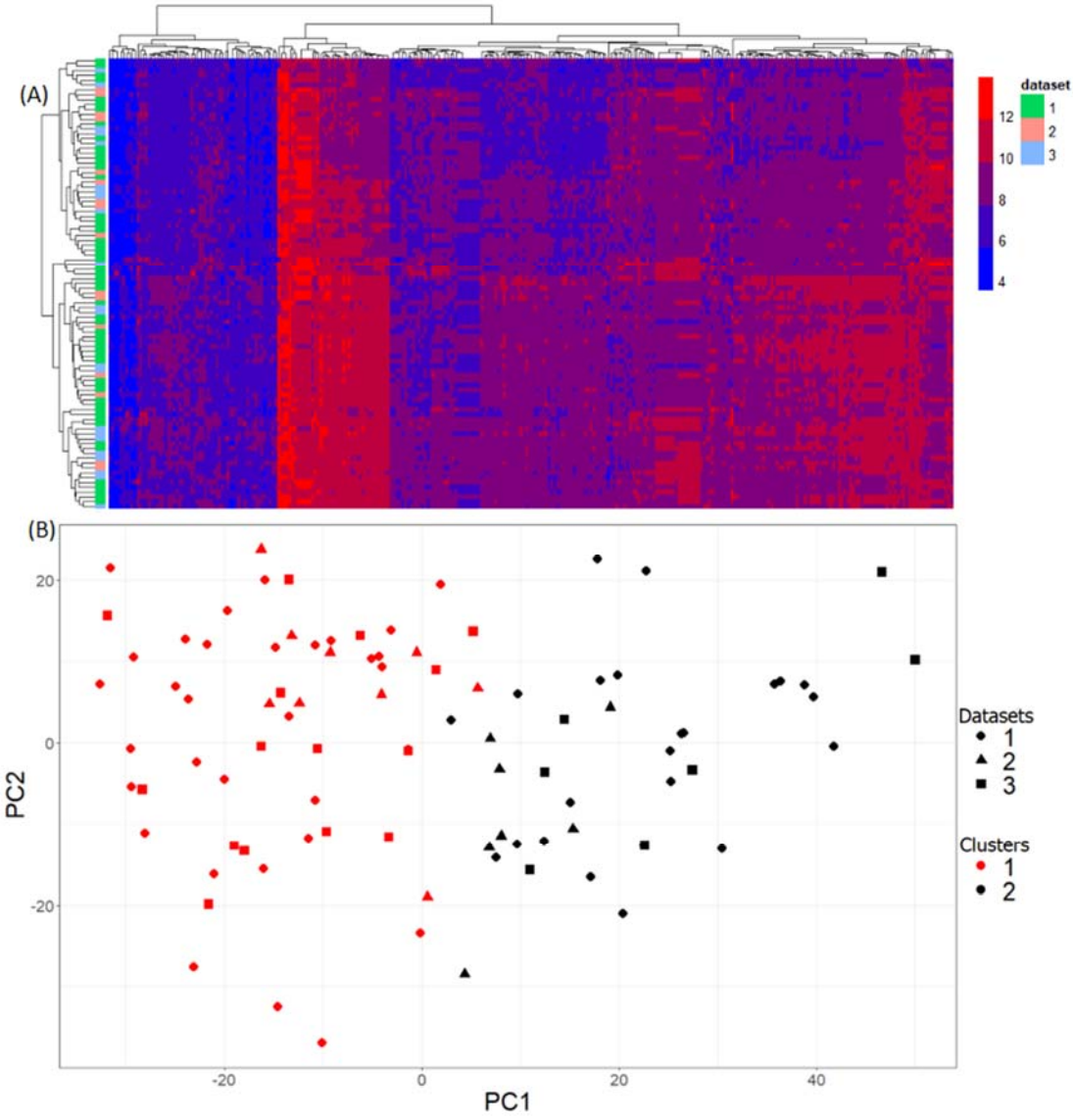

Figure 3: (A) Biclustering analysis of AD transcriptomes in Dataset A (the combined dataset); and (B) PCA shows the two AD-related clusters in Dataset A.

From Fig. 3, one can conclude that the two subgroups of the AD samples are significantly different. From now on, we will refer to the two clusters as subtype 1 and subtype 2 of AD, or "*s1*" and "*s2*" for short. In Additional file Y, we extensively show that if the same procedure is separately applied to Dataset 1 and to Dataset 2+3, the results overlap significantly, which confirms that our "two-subtype model" is not an artifact of dataset combining.

One can ask two questions here:

1. How these two subtypes of AD differ from control samples?
2. What is the difference between *s1* and *s2* at the molecular level?

In all the above analyses, we only considered the AD samples. In order to answer the present questions, it is necessary to include the non-AD samples in the analysis. To this end, we combined AD and non-AD samples of the three datasets, as explained in Materials and Methods. This dataset will be referred to as “Dataset U”. Then, the list of significant DEGs between *s1* and control samples, *s2* and control samples, and also between *s1* and *s2* were computed (LFC>0.4 and FDR<0.05). To our surprise, the GE profile of *s1* samples was found to be very similar to that of the control samples (Figure 4). The results suggest that in subtype 1 of Alzheimer’s disease, the hippocampal transcriptome is not significantly influenced compared to control, despite the appearance of the disease. In contrast, in subtype 2, a considerable number of genes show altered GE patterns compared to control samples.

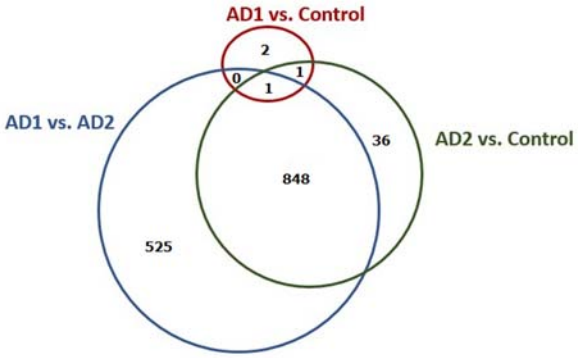

Figure 4: The three sets of statistically significant DEGs (with FDR<0.05 and LFC>|0.4|) in the three comparisons, including AD1 vs. control (4 genes), AD2 vs. control (877 genes), and AD1 vs. AD2 (1849 genes).

All the previously published transcriptome analyses have implicitly assumed AD to be a “homogeneous” disease and determined the DEGs accordingly. Our results, in contrast, suggest that not only AD is heterogeneous, but also it includes a subtype (*i.e.*, *s1*) that has a healthy-like GE profile. Obviously, in such a case, the distribution of samples over the subtypes can influence the detectable DEGs.

We showed that *s1* and *s2* samples do not show significant associations with gender, age and Braak staging. As mentioned in the Introduction, a small group of AD patients does not show atrophy in their hippocampus. Hence, one may presume that this group of patients are those in subtype 1, as this subtype is hardly different from controls samples. However, subtype 1 in our analysis includes a larger number of samples compared to subtype 2, which suggests that the similarity between subtype 1 and control samples must have other reasons, which are yet to be explained.

Pathway enrichment analysis results and comparing them to previous studies

In the next step, using DAVID, gene ontology (GO) and pathway enrichment analyses were performed for the union of two sets of significant DEGs, namely, *“s1* vs. control” and “*s2* vs.

control” comparisons. The results are presented in Additional file 4. The enriched cellular compartment (CC) terms include Cytosol, Extracellular Exosome, Proteasome Complex, etc. The enriched Bioprocess (BP) terms include regulation of cellular amino acid metabolic process, NIK/NF-kappaB signaling, regulation of mRNA stability, etc. The enriched pathways from KEGG database include Oxidative phosphorylation, Alzheimer's disease, etc.

Although poor reproducibility of transcriptome is a known problem in AD studies [18], still one may find a number of commonly enriched GO terms and KEGG pathways in studies focusing on hippocampal transcriptomes of AD subjects. To see if this is the case, we compared our top 20 enriched GO terms/pathways with those reported in four previous studies. The results are shown in Table 4. As expected, there is limited consistency between the enriched sets, which presumably reflects the dataset-to-dataset variability of AD samples.

Table 2: Number of enriched GO terms/pathways which are both reported in our work and in previous studies

Study	No. of reported significant gene sets	Gene set	No. of shared GO terms or pathways
Wu *et al.* [54]	22	GO	3
Peng *et al.* [55]	10	KEGG	6
Hosseinian *et al.* [56]	15	KEGG	3
Lanke *et al.* [30]	26	KEGG	6

We then focused on those 508 genes which show differential GE patterns between *s1* and *s2*, but are not differentially expressed between AD and control samples (Fig. 6). This set includes those genes which are differentially expressed in the two AD subtypes, and additionally, not differentially expressed in AD vs. control samples. We used the DAVID web server to perform GO and pathway enrichment analysis. The results are shown in Additional file 4.

There are certain similarities between the GO terms and KEGG pathways which are enriched both in “AD vs. control” and “subtype 1 vs. subtype 2” comparisons. However, many GO terms and pathways exist which are enriched solely in case of “subtype 1 vs. subtype 2” comparison. One can observe that certain cellular compartment terms related to the cell nucleus (including “nucleus”, “nucleoplasm”, and “nuclear chromosome, telomeric region”), “spliceosomal complex” and “chaperonin-containing T-complex” are exclusively enriched in the latter comparison. Furthermore, “protein polyubiquitination” as a biological process and “ubiquitin mediated proteolysis” as a KEGG pathway is exclusively enriched in the subtype-subtype comparison. These observations provide further evidence about the molecular basis of the difference between the two AD subtypes, which are yet to be studied deeply.

Molecular signatures of the two AD subtypes

Since the two proposed AD subtypes show significantly different GE patterns, we decided to present a “molecular signature” that can be used to distinguish between the two AD subtypes. To achieve this goal, the top 300 DEGs were grouped into 14 clusters based on the similarity

of their gene expression profiles (Figure 5). Then, one gene was selected as the representative of each cluster. The distinctive GE patterns of these 14 genes are shown in Figure 8, and further details on the list of these genes are presented in Additional file 4. We suggest that the set of these 14 genes can be used as a molecular signature whose GE levels can determine whether a sample belongs to subtype 1 or subtype 2 of the AD.

Here, it should be noted that while the present study provides strong evidence about the existence of two subtypes in AD, it can hardly give any insight into the molecular mechanism of AD pathogenesis. Recently, it has been shown that disease-associated changes in AD are highly cell-type specific [57–59]. Therefore, microarray-based transcriptome analysis, which studies “lumped” transcriptomes of multiple cell-types, might be incapable of determining the molecular changes that occur in AD. With the recent advances in the single-cell RNA-seq analysis [60], one might be able to better understand such cell-specific changes across multiple AD subtypes.

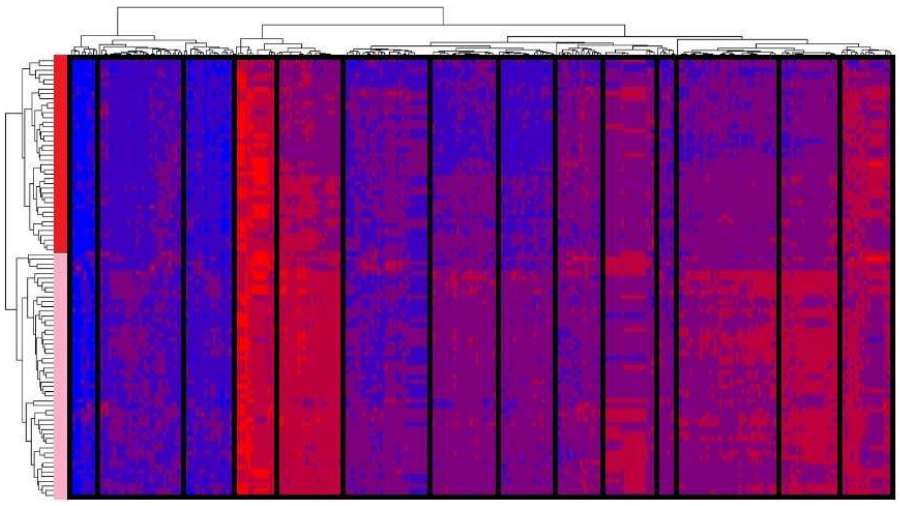

Figure 5: The fourteen gene clusters of the 300 most variable genes between samples of the Combined Dataset AD. The biclustering plot is the same as the plot in Figure 3A.

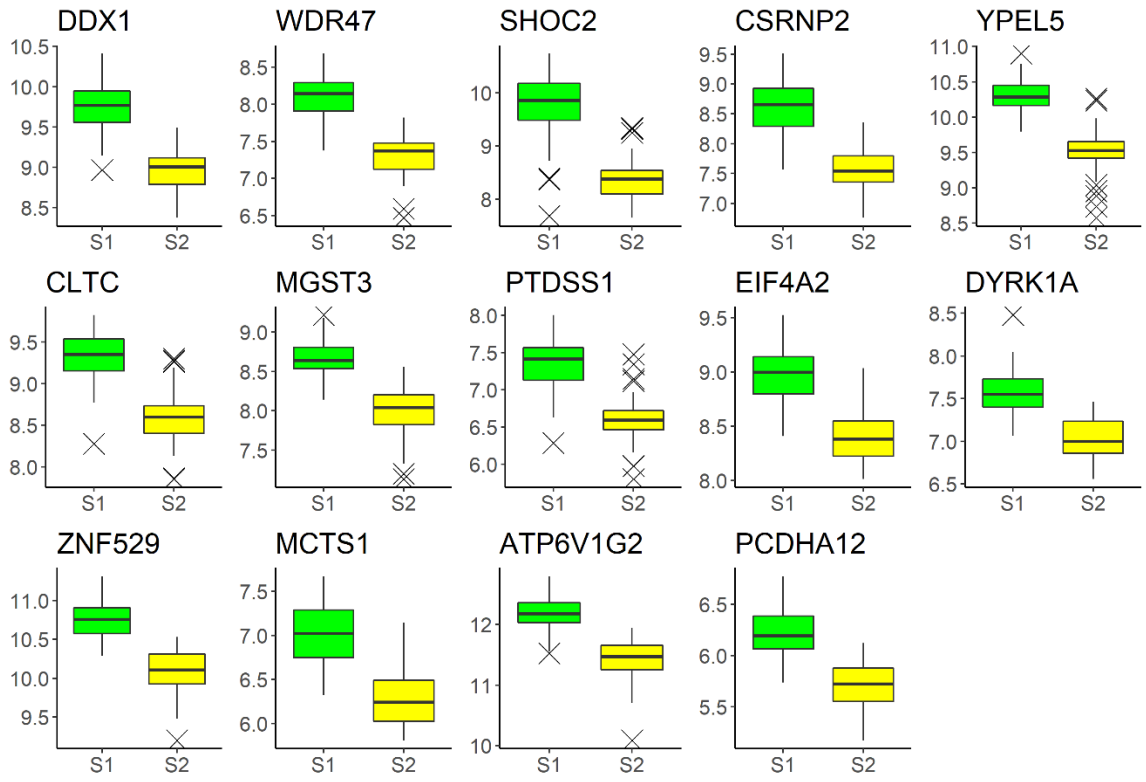

Figure 6: Molecular signature of AD subtypes: comparison of the gene expression profiles of the fourteen selected genes which are differentially expressed across the two AD subtypes. The genes include *ATP6V1G2*, *CLTC*, *CSRNP2*, *DDX1*, *DYRK1A*, *EIF4A2*, *MCTS1*, *MGST3*, *PCDHA12*, *PTDSS1*, *SHOC2*, *WDR47*, *YPEL5* and *ZNF529*.

Conclusion

Previous studies on Alzheimer's disease (AD) have shown heterogeneity in various aspects of the disease. In the present work, we pinpoint that such heterogeneity is also present at the transcriptome level. Based on the meta-analysis of hippocampal transcriptomes, we show that two significantly different subtypes of AD exist. One of these subtypes has a highly similar transcriptomic profile to the transcriptome profiles of healthy (*i.e.*, non-AD) subjects. Furthermore, we nominate signature genes for subtypes based on DEGs and identify altered pathway between subtypes and compared to controls.

This study has several limitations. The samples used in this study (and generally in brain transcriptome studies) are postmortem and the results might be different in alive samples. In addition, despite using multiple hippocampus datasets, the sample size was still limited. Using a larger sample size and analyzing other brain regions in the future could help us shed more light on each subtype's mechanisms and their differences. Low robustness is a common issue in AD transcriptomes studies that we hope to have decreased using a dataset with different microarray platforms and meta-analysis approaches.

Understanding heterogeneity in AD-related transcriptomes can potentially influence attempts to find biomarkers and to assess the validity and/or efficacy of potential treatments. In other

words, this heterogeneity may explain the inconsistencies among transcriptomic studies of AD. We believe that determining AD subtypes based on the GE profiles can be considered as a first step in understanding the complex nature of this disease.

Supplementary information:

***Additional file 1*: Codes of all the statistical analyses.**

***Additional file 2*: Testing the dependence between the two AD subtypes and (a) gender, (b) age, and (c) Braak stage.**

Additional file 3: Validation: The results of our method used separately on Dataset 1 and on Dataset 2+3 show a significant overlap

***Additional file 4*: The results of the gene set enrichment analysis.** Enriched GO terms or KEGG pathways in both the “AD vs. control” and the “subtype 1 vs. subtype 2” comparisons are highlighted.

***Additional file 5*: List of the 14 genes identified in this study as the molecular signature to distinguish the two subtypes of Alzheimer’s disease.**

References

1. Mietlicki-Baase EG. Amylin in Alzheimer’s disease: Pathological peptide or potential treatment? Neuropharmacology. 2018,136:287–97. doi:10.1016/j.neuropharm.2017.12.016.
2. Khayer N, Marashi SA, Mirzaie M, Goshadrou F. Three-way interaction model to trace the mechanisms involved in Alzheimer’s disease transgenic mice. PLoS One. 2017,12:1–18.
3. 2019 Alzheimer’s disease facts and figures. Alzheimer’s Dement. 2019,15:321–87. doi:10.1016/j.jalz.2019.01.010.
4. Grady CL, Haxby J V., Horwitz B, Sundaram M, Berg G, Schapiro M, et al. Longitudinal study of the early neuropsychological and cerebral metabolic changes in dementia of the Alzheimer type. J Clin Exp Neuropsychol. 1988,10:576–96. doi:10.1080/01688638808402796.
5. Galton CJ. Atypical and typical presentations of Alzheimer’s disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. Brain. 2000,123:484–98. doi:10.1093/brain/123.3.484.
6. Niu H, Álvarez-Álvarez I, Guillén-Grima F, Aguinaga-Ontoso I. Prevalence and incidence of Alzheimer’s disease in Europe: A meta-analysis. Neurología (English Ed.) 2017,32:523–32. doi:10.1016/j.nrleng.2016.02.009.
7. Lam B, Masellis M, Freedman M, Stuss DT, Black SE. Clinical, imaging, and pathological heterogeneity of the Alzheimer’s disease syndrome. Alzheimers Res Ther. 2013,5:1. doi:10.1186/alzrt155.
8. Shoji M, Harigaya Y. Biochemical heterogeneity of Alzheimer’s disease. Neuropathology. 1998,18:116–20. doi:10.1111/j.1440-1789.1998.tb00088.x.
9. Au R, Piers RJ, Lancashire L. Back to the future: Alzheimer’s disease heterogeneity revisited. Alzheimer’s Dement Diagnosis, Assess Dis Monit. 2015,1:368–70. doi:10.1016/j.dadm.2015.05.006.
10. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991,82:239–59. doi:10.1007/BF00308809.
11. Braak H, Braak E. Staging of alzheimer’s disease-related neurofibrillary changes. Neurobiol Aging. 1995,16:271–8. doi:10.1016/0197-4580(95)00021-6.
12. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined

subtypes of Alzheimer's disease with distinct clinical characteristics: A retrospective study. Lancet Neurol. 2011,10:785–96. doi:10.1016/S1474-4422(11)70156-9.
13. Ferreira D, Pereira JB, Volpe G, Westman E. Subtypes of Alzheimer's Disease Display Distinct Network Abnormalities Extending Beyond Their Pattern of Brain Atrophy. Front Neurol. 2019,10:524. doi:10.3389/fneur.2019.00524.
14. Whitwell JL, Dickson DW, Murray ME, Weigand SD, Tosakulwong N, Senjem ML, et al. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: A case-control study. Lancet Neurol. 2012,11:868–77.
15. Ferreira D, Verhagen C, Hernández-Cabrera JA, Cavallin L, Guo C-J, Ekman U, et al. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. Sci Rep. 2017,7:46263. doi:10.1038/srep46263.
16. Koedam ELGE, Lauffer V, van der Vlies AE, van der Flier WM, Scheltens P, Pijnenburg YAL. Early-Versus Late-Onset Alzheimer's Disease: More than Age Alone. J Alzheimer's Dis. 2010,19:1401–8. doi:10.3233/JAD-2010-1337.
17. Dong A, Toledo JB, Honnorat N, Doshi J, Varol E, Sotiras A, et al. Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. Brain. 2017,140:735–47.
18. Scheltens NME, Galindo-Garre F, Pijnenburg YAL, van der Vlies AE, Smits LL, Koene T, et al. The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. J Neurol Neurosurg Psychiatry. 2016,87:235–43. doi:10.1136/jnnp-2014-309582.
19. Annese A, Manzari C, Lionetti C, Picardi E, Horner DS, Chiara M, et al. Whole transcriptome profiling of Late-Onset Alzheimer's Disease patients provides insights into the molecular changes involved in the disease. Sci Rep. 2018,8:4282. doi:10.1038/s41598-018-22701-2.
20. Verheijen J, Sleegers K. Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics. Trends Genet. 2018,34:434–47. doi:10.1016/j.tig.2018.02.007.
21. Hadar A, Gurwitz D. Peripheral transcriptomic biomarkers for early detection of sporadic Alzheimer disease? Dialogues Clin Neurosci. 2018,20:293–300. doi:10.31887/DCNS.2018.20.4/dgurwitz.
22. Berchtold NC, Coleman PD, Cribbs DH, Rogers J, Gillen DL, Cotman CW. Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. Neurobiol Aging. 2013,34:1653–61. doi:10.1016/j.neurobiolaging.2012.11.024.
23. Hokama M, Oka S, Leon J, Ninomiya T, Honda H, Sasaki K, et al. Altered Expression of Diabetes-Related Genes in Alzheimer's Disease Brains: The Hisayama Study. Cereb Cortex. 2014,24:2476–88. doi:10.1093/cercor/bht101.
24. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. PLoS One. 2011,6:e16266. doi:10.1371/journal.pone.0016266.
25. Zhao L, Lee VHF, Ng MK, Yan H, Bijlsma MF. Molecular subtyping of cancer: current status and moving toward clinical applications. Brief Bioinform. 2019,20:572–84. doi:10.1093/bib/bby026.
26. Wang W, Kandimalla R, Huang H, Zhu L, Li Y, Gao F, et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. Semin Cancer Biol. 2019,55:37–52. doi:10.1016/j.semcancer.2018.05.002.
27. Bagyinszky E, Giau V Van, An SA. Transcriptomics in Alzheimer's Disease: Aspects and Challenges. Int J Mol Sci. 2020,21:3517. doi:10.3390/ijms21103517.
28. Li Y, Wu Z, Jin Y, Wu A, Cao M, Sun K, et al. Analysis of hippocampal gene expression profile of Alzheimer's disease model rats using genome chip bioinformatics. Neural Regen Res. 2012,7:332–40. doi:10.3969/j.issn.1673-5374.2012.05.002.
29. Dharshini SAP, Taguchi Y -h., Gromiha MM. Investigating the energy crisis in Alzheimer disease using transcriptome study. Sci Rep. 2019,9:18509. doi:10.1038/s41598-019-54782-y.
30. Lanke V, Moolamalla STR, Roy D, Vinod PK. Integrative Analysis of Hippocampus Gene Expression Profiles Identifies Network Alterations in Aging and Alzheimer's Disease. Front Aging Neurosci. 2018;10:153. doi:10.3389/fnagi.2018.00153.
31. Padurariu M, Ciobica A, Mavroudis I, Fotiou D, Baloyannis S. Hippocampal neuronal loss in the CA1 and CA3 areas of Alzheimer's disease patients. Psychiatr Danub. 2012,24:152–8. http://www.ncbi.nlm.nih.gov/pubmed/22706413.
32. Mufson EJ, Mahady L, Waters D, Counts SE, Perez SE, DeKosky ST, et al. Hippocampal plasticity during the progression of Alzheimer's disease. Neuroscience. 2015,309:51–67. doi:10.1016/j.neuroscience.2015.03.006.
33. Wilson IA, Gallagher M, Eichenbaum H, Tanila H. Neurocognitive aging: prior memories hinder new

hippocampal encoding. Trends Neurosci. 2006,29:662–70. doi:10.1016/j.tins.2006.10.002.
34. Setti SE, Hunsberger HC, Reed MN. Alterations in hippocampal activity and Alzheimer's disease. Transl Issues Psychol Sci. 2017,3:348–56. doi:10.1037/tps0000124.
35. Neff RA, Wang M, Vatansever S, Guo L, Ming C, Wang Q, et al. Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. Sci Adv. 2021,7:1–18.
36. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005,21:171–8. doi:10.1093/bioinformatics/bth469.
37. Tan PK. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 2003,31:5676–84. doi:10.1093/nar/gkg763.
38. Irigoyen A, Jimenez-Luna C, Benavides M, Caba O, Gallego J, Ortuño FM, et al. Integrative multi-platform meta-analysis of gene expression profiles in pancreatic ductal adenocarcinoma patients for identifying novel diagnostic biomarkers. PLoS One. 2018,13:e0194844. doi:10.1371/journal.pone.0194844.
39. Patel H, Dobson RJB, Newhouse SJ. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. J Alzheimer's Dis. 2019,68:1635–56. doi:10.3233/JAD-181085.
40. Su L, Chen S, Zheng C, Wei H, Song X. Meta-Analysis of Gene Expression and Identification of Biological Regulatory Mechanisms in Alzheimer's Disease. Front Neurosci. 2019,13. doi:10.3389/fnins.2019.00633.
41. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012,41:D991–5. doi:10.1093/nar/gks1193.
42. Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Med. 2016,8:104. doi:10.1186/s13073-016-0355-3.
43. Miller JA, Woltjer RL, Goodenbour JM, Horvath S, Geschwind DH. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome Med. 2013,5:48. doi:10.1186/gm452.
44. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci. 2004,101:2173–8. doi:10.1073/pnas.0308512100.
45. Turnbull AK, Kitchen RR, Larionov AA, Renshaw L, Dixon JM, Sims AH. Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. BMC Med Genomics. 2012,5:35. doi:10.1186/1755-8794-5-35.
46. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012,6:80–92. doi:10.4161/fly.19695.
47. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. BMC Genomics. 2009,10:22. doi:10.1186/1471-2164-10-22.
48. Ben B. preprocessCore: A collection of pre-processing functions. 2020. https://github.com/bmbolstad/preprocessCore.
49. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012,28:882–3. doi:10.1093/bioinformatics/bts034.
50. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. 2019.
51. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987,20:53–65. doi:10.1016/0377-0427(87)90125-7.
52. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015,43:e47–e47. doi:10.1093/nar/gkv007.
53. Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, et al. Extracting Biological Meaning from Large Gene Lists with DAVID. Curr Protoc Bioinforma. 2009,27:13.11.1-13.11.13. doi:10.1002/0471250953.bi1311s27.
54. Wu M, Fang K, Wang W, Lin W, Guo L, Wang J. Identification of key genes and pathways for Alzheimer's disease via combined analysis of genome-wide expression profiling in the hippocampus. Biophys Reports. 2019,5:98–109. doi:10.1007/s41048-019-0086-2.
55. Peng Y-S, Tang C-W, Peng Y-Y, Chang H, Chen C-L, Guo S-L, et al. Comparative functional genomic analysis of Alzheimer's affected and naturally aging brains. PeerJ. 2020,8:e8682. doi:10.7717/peerj.8682.
56. Hosseinian S, Arefian E, Rakhsh-Khorshid H, Eivani M, Rezayof A, Pezeshk H, et al. A meta-analysis of gene expression data highlights synaptic dysfunction in the hippocampus of brains with Alzheimer's disease. Sci Rep. 2020,10:8384. doi:10.1038/s41598-020-64452-z.
57. Johnson TS, Xiang S, Dong T, Huang Z, Cheng M, Wang T, et al. Combinatorial analyses reveal cellular

composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer's Disease. Sci Rep. 2021,11:353. doi:10.1038/s41598-020-79740-x.
58. Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci. 2019,22:2087–97. doi:10.1038/s41593-019-0539-4.
59. Lau S, Cao H, Fu AKY, Ip NY. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. Proc Natl Acad Sci. 2020,117:25800–9. doi:10.1073/pnas.2008762117.
60. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. 2019,10:1–13.

Declarations:

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All gene expression datasets are obtained from GEO (see Table 1) and are publicly available. The codes are available in *Additional file 1*.

Funding
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author Contributions
SAM and HM presented the original idea. All the computational analyses were performed by AN and HM. All authors were involved in designing the computational experiments and interpreting the results. The manuscript is drafted by AN and SAM. All authors read and approved the final manuscript.

Acknowledgments:
Not applicable.

Code availability: All data analyses were performed in the R environment. The codes are available in *Additional file 1*.

Figures

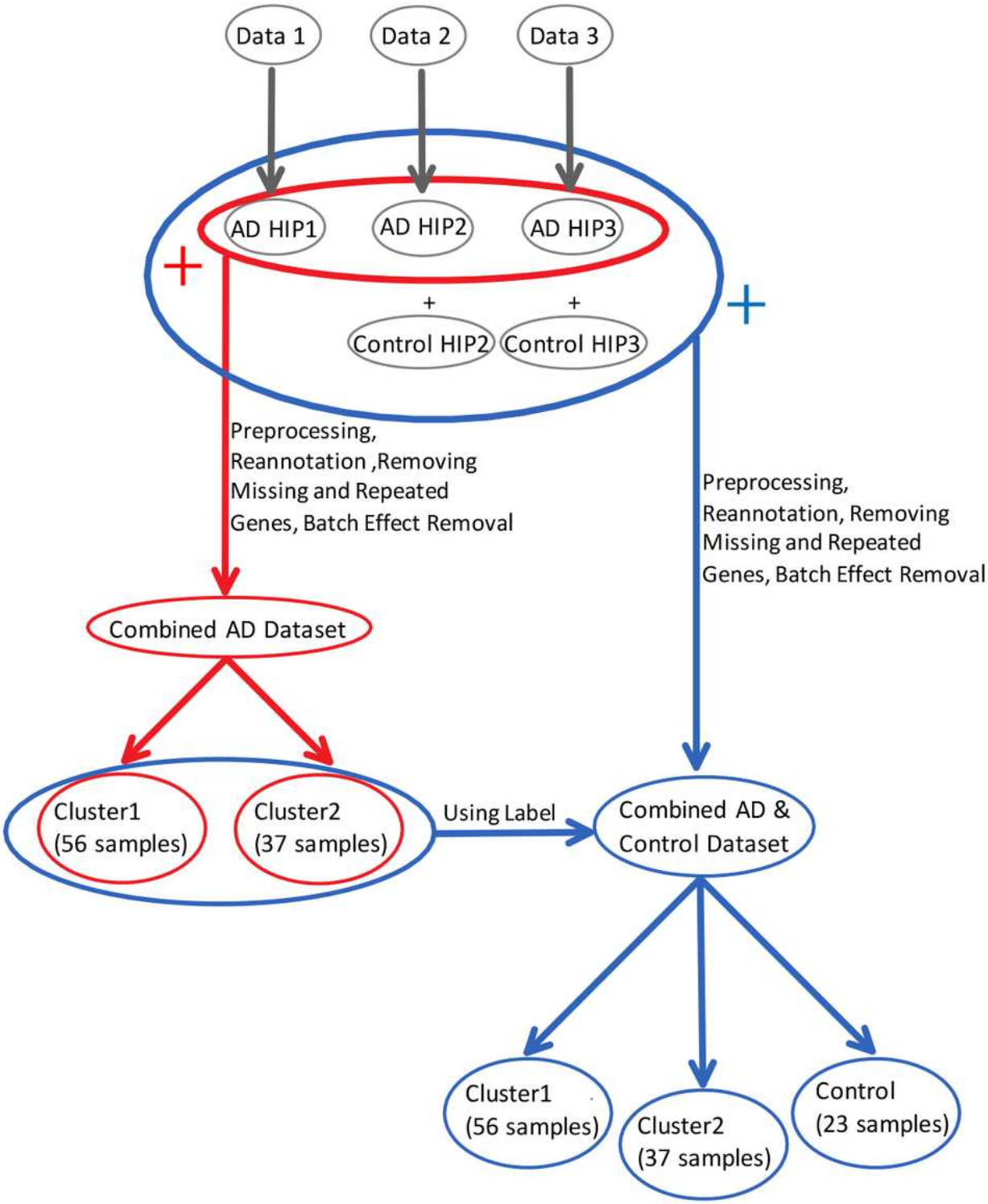

Figure 1

Schematic representation of the data processing for normalizing and combining the three datasets. The combining protocol is based on previously published methods [38, 45]. Briefly, Affymetrix data were preprocessed by 'rma' function in 'affy' package [46] and re-annotated by biomaRt [47]. The Illumina data

were “log2 transformed”, and then, normalized (by quantile normalization approach from package 'preprocessCore' [48]) and re-annotated by biomaRt. After removing gene redundancies, for the combined dataset, batch effect removal was performed by ComBat from ‘sva’ package [49]. (Red steps: combining only AD samples to find AD clusters; Blue steps: combining AD and controls samples and applying the cluster labels obtained in the “red steps”).

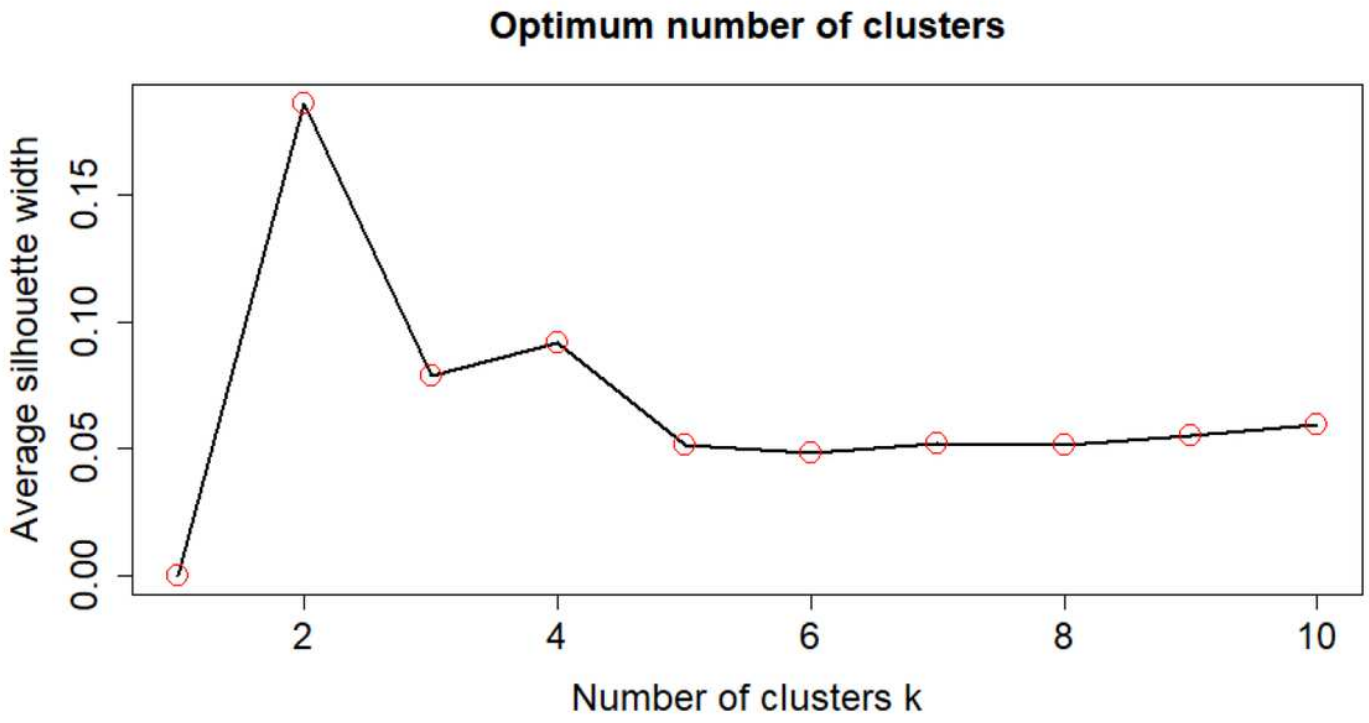

Figure 2

Average silhouette width for partitioning Dataset 1 into k=2,3,…,10 clusters

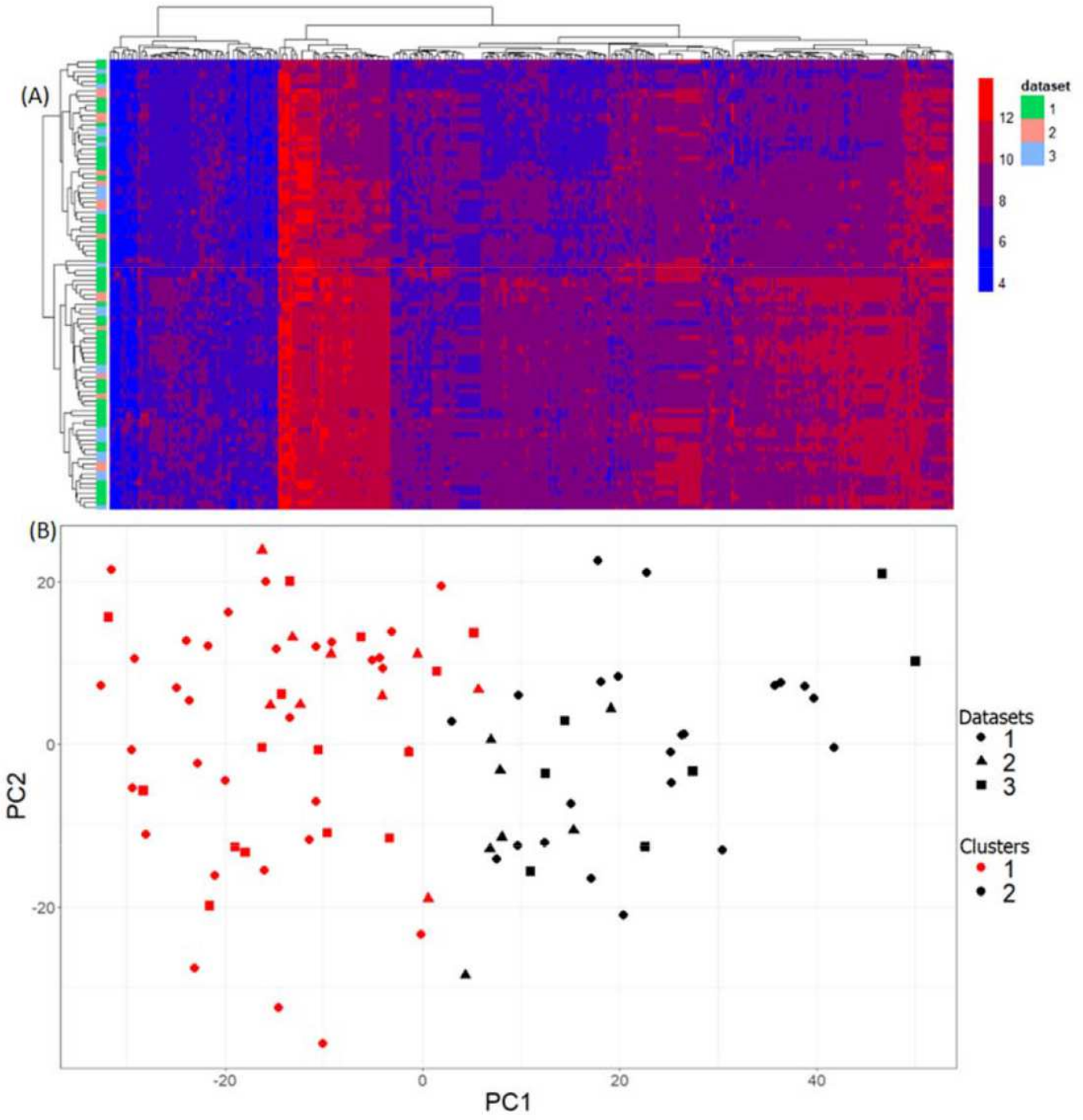

Figure 3

(A) Biclustering analysis of AD transcriptomes in Dataset A (the combined dataset); and (B) PCA shows the two ADrelated clusters in Dataset A.

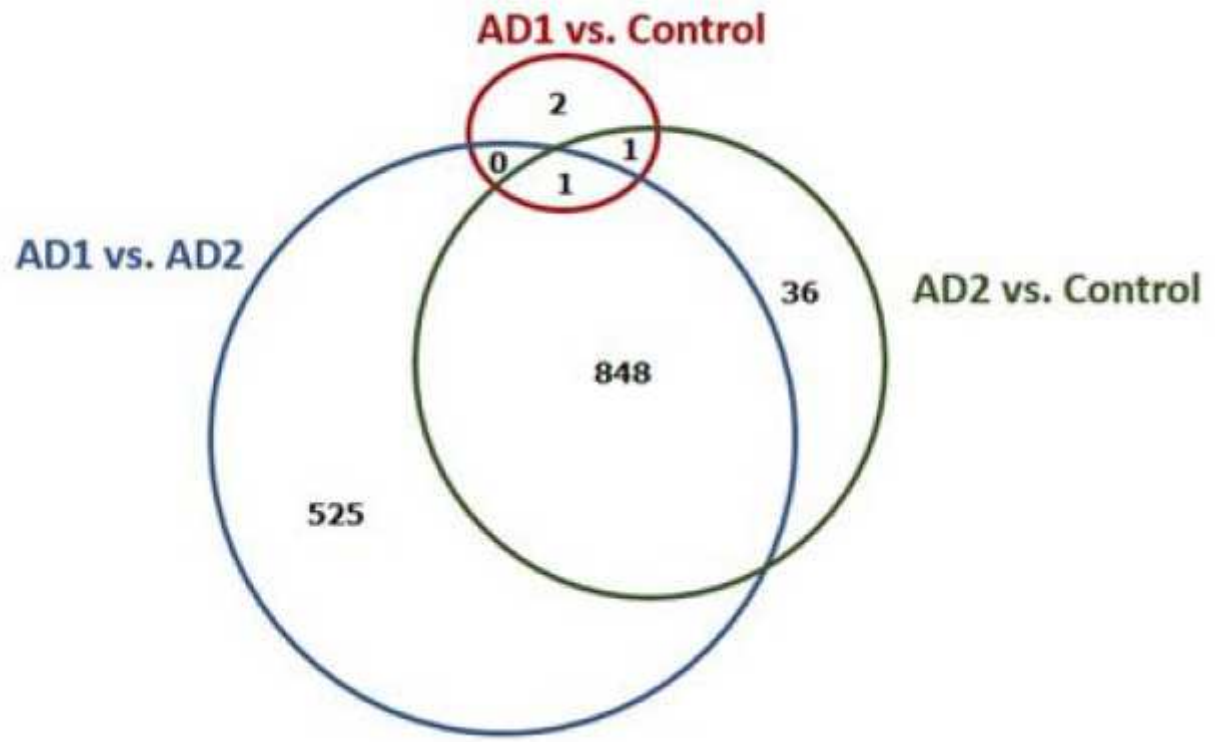

Figure 4

The three sets of statistically significant DEGs (with FDR<0.05 and LFC>|0.4|) in the three comparisons, including AD1 vs. control (4 genes), AD2 vs. control (877 genes), and AD1 vs. AD2 (1849 genes).

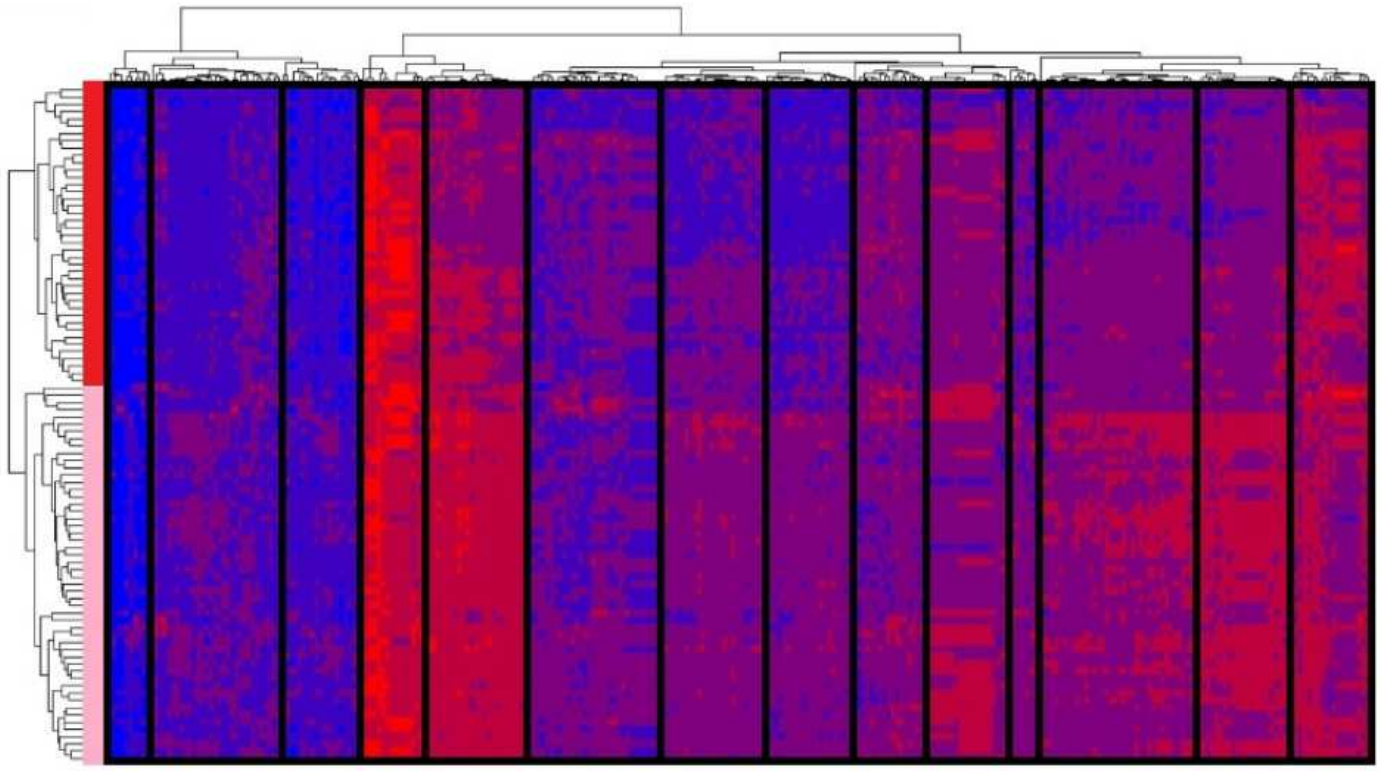

Figure 5

The fourteen gene clusters of the 300 most variable genes between samples of the Combined Dataset AD. The biclustering plot is the same as the plot in Figure 3A.

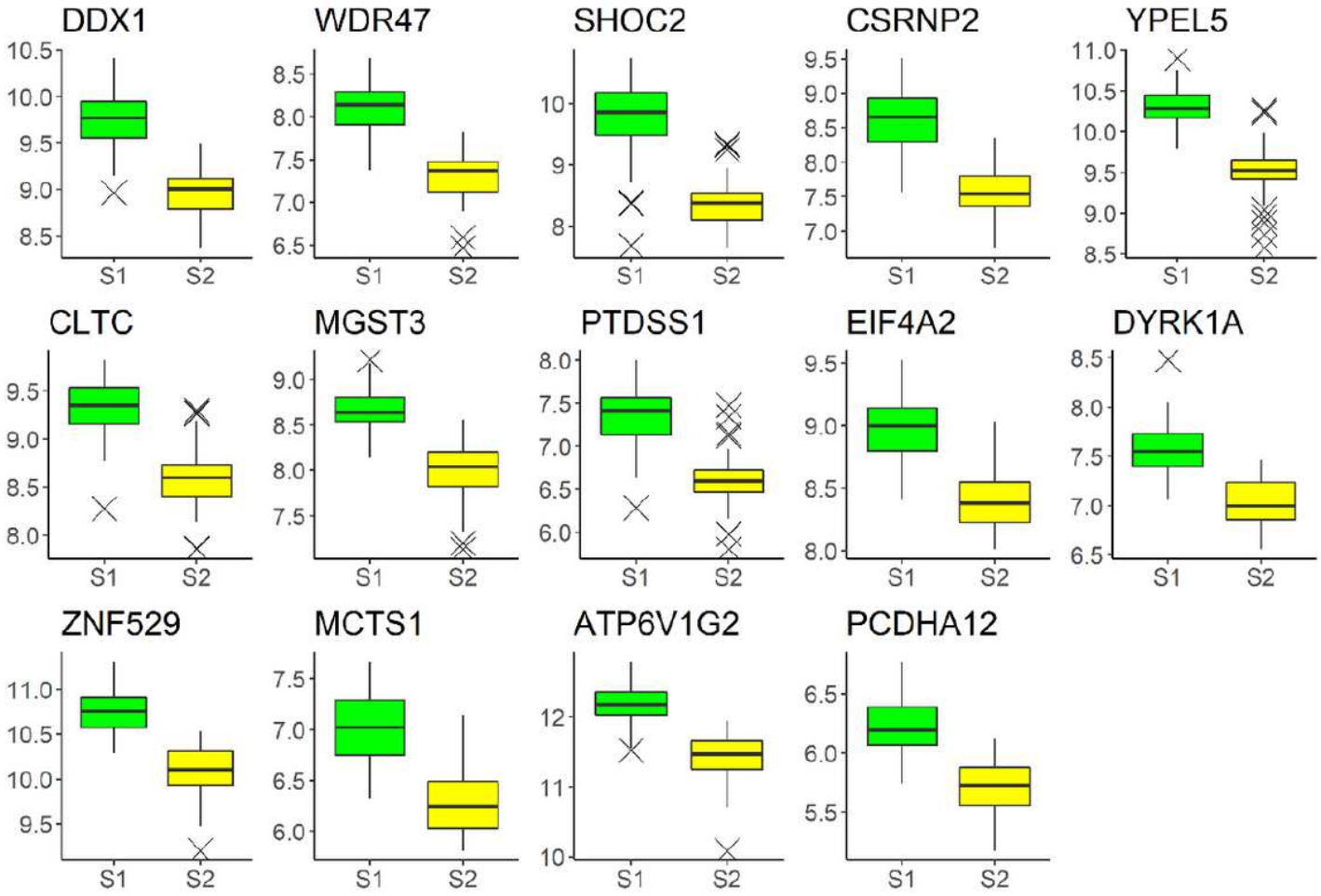

Figure 6

Molecular signature of AD subtypes: comparison of the gene expression profiles of the fourteen selected genes which are differentially expressed across the two AD subtypes. The genes include ATP6V1G2, CLTC, CSRNP2, DDX1, DYRK1A, EIF4A2, MCTS1, MGST3, PCDHA12, PTDSS1, SHOC2, WDR47, YPEL5 and ZNF529.

Supplementary Files

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

- Aditionalfile1.rar
- Additionalfile2.csv
- Additionalfile3.pdf
- Additionalfile4.xlsx
- Additionalfile5.xlsx