

UCSF

UC San Francisco Previously Published Works

Title

Correlation between Alzheimer's disease and type 2 diabetes using non-negative matrix factorization

Permalink

<https://escholarship.org/uc/item/9mh547mh>

Journal

Scientific Reports, 11(1)

ISSN

2045-2322

Authors

Chung, Yeonwoo

Lee, Hyunju

Publication Date

2021

DOI

10.1038/s41598-021-94048-0

Peer reviewed



OPEN

Correlation between Alzheimer's disease and type 2 diabetes using non-negative matrix factorization

Yeonwoo Chung¹, Hyunju Lee¹✉ & the Alzheimer's Disease Neuroimaging Initiative*

Alzheimer's disease (AD) is a complex and heterogeneous disease that can be affected by various genetic factors. Although the cause of AD is not yet known and there is no treatment to cure this disease, its progression can be delayed. AD has recently been recognized as a brain-specific type of diabetes called type 3 diabetes. Several studies have shown that people with type 2 diabetes (T2D) have a higher risk of developing AD. Therefore, it is important to identify subgroups of patients with AD that may be more likely to be associated with T2D. We here describe a new approach to identify the correlation between AD and T2D at the genetic level. Subgroups of AD and T2D were each generated using a non-negative matrix factorization (NMF) approach, which generated clusters containing subsets of genes and samples. In the gene cluster that was generated by conventional gene clustering method from NMF, we selected genes with significant differences in the corresponding sample cluster by Kruskal–Wallis and Dunn-test. Subsequently, we extracted differentially expressed gene (DEG) subgroups, and candidate genes with the same regulation direction can be extracted at the intersection of two disease DEG subgroups. Finally, we identified 241 candidate genes that represent common features related to both AD and T2D, and based on pathway analysis we propose that these genes play a role in the common pathological features of AD and T2D. Moreover, in the prediction of AD using logistic regression analysis with an independent AD dataset, the candidate genes obtained better prediction performance than DEGs. In conclusion, our study revealed a subgroup of patients with AD that are associated with T2D and candidate genes associated between AD and T2D, which can help in providing personalized and suitable treatments.

The number of people worldwide suffering from Alzheimer's disease (AD) has been steadily increasing in recent decades¹. AD is an irreversible disease that slowly and progressively destroys the brain. Specifically, AD affects memory dysfunction, representing a major cause of dementia in the aging population^{2,3}. However, AD is not a normal component of aging, but is instead a complex disease entity, and the detailed pathogenic mechanisms underlying the disease remain unclear. It has been generally recognized that accumulation of plaques (beta-amyloid) and tangles (tau) is the leading cause of AD⁴, and the prime suspect contributing to the associated neuronal destruction^{5,6}. Although the specific causes of beta-amyloid and tau accumulation are unknown, this pathogenic event is considered to be the result of various interacting genetic and environmental factors⁷. Therefore, it is important to address the complexity of AD by detecting the underlying characteristics.

One approach to disentangle a complex disease is gene expression analysis, including the identification of potential candidate genes or comparing expression values for specific AD-related genes. Indeed, several studies have discovered AD-related genes and mechanisms using genome-wide analyses^{8–11}. In particular, AD analyses using blood samples from patients have received considerable attention as a novel method of diagnosis given the advantages of the non-invasive nature and less expensive process compared to traditional analyses using brain tissue or imaging. In fact, several differentially expressed proteins in the AD brain have been identified in the blood of AD patients¹². Bu et al.¹³ showed that the amyloid-beta protein is produced not only in the brain, but also in the peripheral tissues of AD patients. Therefore, many studies have focused on machine-learning and statistical methods to improve early detection and for the identification of candidate genes that can serve as targets for AD treatment based on data obtained from peripheral samples^{14–17}.

Moreover, recent studies have indicated an association of AD with other diseases^{18–20}. Several genetic factors and pathophysiological mechanisms associated with AD are also shared with other AD-related diseases²¹. Therefore, identifying the relationships between AD and AD-related diseases can help to address the complexity of AD

¹School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Korea. *List of authors and their affiliations appear at the end of the paper. ✉email: hyunjulee@gist.ac.kr

and bring us a step closer to realizing the personalized treatment of AD by reducing risk factors from comorbidities and giving a chance to treat common dysregulated pathways between AD and its related diseases^{22,23}. Narayanan et al.²⁴ investigated co-regulated genes in AD and Huntington's disease, and identified common differentially co-expressed subnetworks in the two neurodegenerative diseases. In particular, type 2 diabetes (T2D) is highly associated with AD, and epidemiological studies have shown that patients with T2D have a higher risk of developing AD²⁵. One of the hallmark pathophysiological features in T2D patients is the deposition of amyloid converted from islet amyloid peptide (IAPP)²⁶. Overproduction of IAPP secreted by pancreatic beta-cells may cause beta cell loss in T2D, and there are evidences that intracellular toxic amyloid peptide oligomers are associated with AD²⁷. Given that several lines of evidence indicate a link between AD and T2D, AD can be considered as a brain-specific type of diabetes that has been dubbed “type 3 diabetes”, and inflammation, insulin resistance, and mitochondrial dysfunction were considered as common pathogenesis of two diseases^{28–30}.

Accordingly, the aim of this study was to detect the common and distinct characteristics of AD and T2D using gene expression datasets. Because not all AD patients show an association with T2D, we extracted genes with significantly different expression profiles in AD and T2D patients separately using computational and statistical methods, and then defined subgroups of AD and T2D compared to their respective controls. Based on this analysis, we identified common genes with the same regulation directions (up or downregulated in the disease) in each pair of patient subgroups. Through this approach, we identified candidate genes among the common genes with significant differences in expression levels in the subgroups of the two diseases, which were tested as potential biomarkers for diagnostic or prognostic prediction using an independent Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset³¹. Figure 1 illustrates the procedure for relation extraction between two diseases.

Methods

Data description. The mRNA expression datasets for AD and T2D were downloaded from the Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). For AD analysis, we used integrated data on peripheral blood gene expression profiles from the GSE63060 and GSE63061 datasets using an R software³², which were generated on the Illumina HumanHT-12 v3.0 Expression BeadChip and the Illumina HumanHT-12 v4.0 Expression BeadChip, respectively. The GSE63060 dataset contains data for 145 AD patients and 104 control samples, and the GSE63061 dataset contains data for 140 AD patients and 135 control samples. For T2D, we used gene expression data (GSE78721) including 68 T2D patients and 62 healthy controls from adipocytes and infiltration macrophages because it was the largest T2D dataset in GEO, and adipose-derived transcription signature is associated with T2D^{33,34}. Gene expression data were generated using the Affymetrix PrimeView Human Gene Expression Array.

For each gene expression data point from the GEO datasets, the probe ID was converted into the Entrez Gene ID using information from the platforms of each AD and T2D datasets (e.g., GPL6947, GPL10558, and GPL15207). Gene expression levels of the probe that did not match the Entrez Gene ID were removed. Among all the assigned Entrez Gene IDs, protein-coding genes were selected using database Homo_sapiens.GRCh38.94 (http://asia.ensembl.org/Homo_sapiens/Info/Index), where Ensembl IDs were converted to the Entrez Gene IDs using the “biomaRt” package in R software. For duplicated Entrez Gene IDs, the expression values of the same Entrez Gene ID were merged into the mean value. To analyze the AD data, we merged the two datasets (GSE63060 and GSE63061). The number of Entrez Gene IDs selected by the protein-coding gene database in each AD dataset was 16,730 and 14,957, respectively. We selected 14,134 common genes from the two datasets and used the “removeBatchEffect” function in the R package “limma”. As a result, we obtained the expression data of 14,134 genes from the 285 AD patients and 239 control samples. In addition, we normalized each data point in the patient expression data by applying a $\log_2(\text{fold change})$ conversion as follows:

$$y_{ij} = \log_2 \frac{\text{patient}_{ij}}{\text{normal}_i}$$

where patient_{ij} is the expression value of the gene i of the j th patient and normal_i is the mean expression value of the gene i in the normal control samples.

Decomposition of gene expression data sets. Most gene expression datasets contain information on thousands of genes, which is relatively large compared to the number of samples; therefore, several studies have applied dimensionality reduction methods to reduce the matrix dimension^{35,36}. The non-negative matrix factorization (NMF) method has been widely used to reduce the dimension of an input matrix by decomposing a non-negative input matrix into two matrices³⁷. Assuming an input matrix A consisting of the expression data of N genes and M samples, NMF produces the matrices W and H of size $N \times k$ and $k \times M$, respectively, in which the parameter k indicates the number of clusters desired in the input data. The NMF algorithm is a multiplication update algorithm that multiplies W and H to obtain the input A during iteration until convergence. After convergence, the matrices W and H are used to bi-cluster genes and samples^{38,39}. Each row of W ($\text{gene} \times k$) and each column of H ($k \times \text{sample}$) can be represented by a positive linear combination of k . The element w_{ij} in the W matrix is the coefficient of gene i and cluster j ($1 \sim k$), and the element h_{ij} in the H matrix is the coefficient of cluster i ($1 \sim k$) and sample j (Fig. 1b).

Determination of the optimal number of clusters according to the rank k . To select the optimal number of meaningful clusters that correctly divide the input data, the cophenetic correlation coefficient should be taken into consideration³⁷. NMF updates the W_i and H_i matrices of each i th iteration until convergence. Following this, a connectivity matrix C_i (with size $\text{sample} \times \text{sample}$) is defined by each sample assignment of the

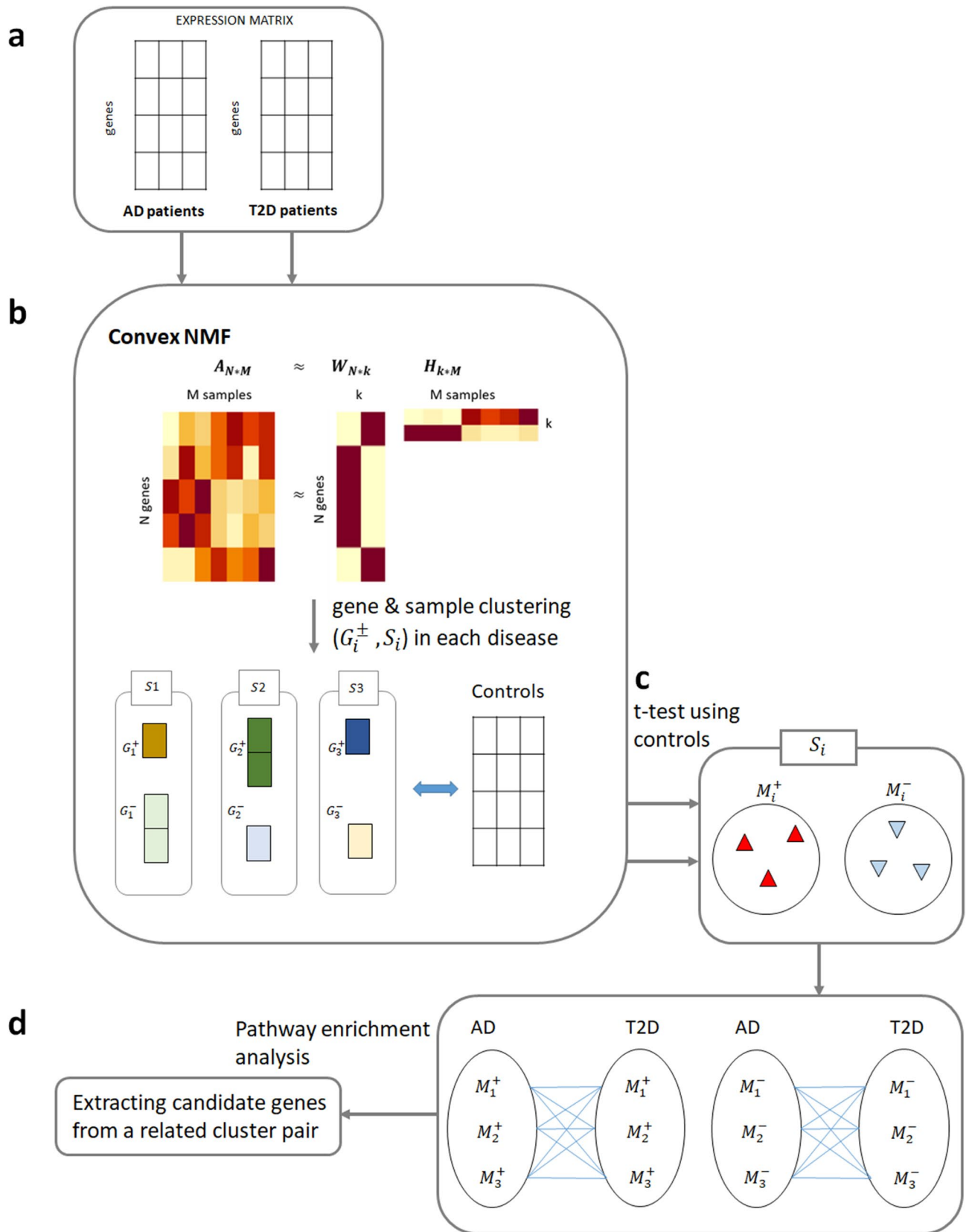


Figure 1. Procedure for relation extraction between two diseases. (a) Gene expression data of AD and T2D patients are given. (b) The convex non-negative matrix factorization (NMF) is used to decompose the input expression matrices. Gene and sample clusters are obtained from the NMF decomposed matrices. (c) DEG genes are assigned to clusters. (d) Common candidate genes are extracted from a related cluster pair between two diseases. (b) is generated by the R software (R version 3.6.1, <https://www.r-project.org/>) using an example dataset.

i th iteration from H_i by selecting the maximum index of each column. The elements of the connectivity matrix c_{ij} are filled with 1 s if the samples i and j are assigned to the same cluster and are filled with 0 s otherwise. The average of all connectivity matrices represents the consensus matrix \bar{C} , which is the probability that the samples i and j belong together during iterations. The cophenetic correlation coefficient is then calculated as the Pearson correlation coefficient between $I - \bar{C}$ and the distance between samples in a hierarchical clustering of \bar{C} . The cophenetic correlation coefficient indicates the dispersion of the sample assignment, which refers to how consistently samples with similar gene expression profiles belong together during iterations. Therefore, the rank k with the highest cophenetic correlation indicates the optimal capacity of the model.

However, NMF can only process non-negative ranges of entries in the input matrix A , and the output matrices W and H also have non-negative ranges. Therefore, to analyze the $\log_2(\text{fold change})$ expression dataset of each disease, we needed to select a model that can cover both positive and negative values. If the range of the input matrix A is \pm , then the matrix A_{\pm} is decomposed such that $A_{\pm} \approx W_{\pm} \times H_{\pm}$. In convex NMF, the basis vectors of the W_{\pm} matrix are considered to be convex combinations of the input matrix A_{\pm} (i.e., $A_{\pm} \approx W_{\pm} \times H_{\pm} \approx X_{\pm} \times F_{\pm} \times H_{\pm}$ ⁴⁰) and there is an advantage in that the factors of F_{\pm} and H_{\pm} are sparse. Therefore, we apply the convex NMF to obtain the W and H matrices. $X_{\pm} \times F_{\pm}$ from convex NMF is treated as a factor of W in the NMF used for gene clustering, and then F_{+} and H_{+} are updated alternatively as follows:

$$H_{ik} \leftarrow H_{ik} \sqrt{\frac{[(X^T X) + F]_{ik} + [HF^T (X^T X) - F]_{ik}}{[(X^T X) - F]_{ik} + [HF^T (X^T X) + F]_{ik}}}$$

$$F_{ik} \leftarrow F_{ik} \sqrt{\frac{[(X^T X) + H]_{ik} + [(X^T X) - FH^T H]_{ik}}{[(X^T X) - H]_{ik} + [(X^T X) + FH^T H]_{ik}}}$$

Gene and sample clusters. The basic method of gene and sample clustering in NMF using the factors W and H is the “Max” method, which selects the cluster with the highest coefficient³⁷. In general, a gene is assigned to the cluster with the highest coefficient in each gene row in the W matrix, and a sample is assigned to cluster S_i when the i th coefficient is the highest coefficient in each sample column in the H matrix. Accordingly, the gene cluster obtained by the “Max” gene clustering method is a group of genes with relatively upregulated expression in the sample cluster S_i compared to other sample clusters. Conversely, to consider the gene cluster with relatively downregulated expression, a gene is assigned to the cluster with the minimum coefficient using the “Min” gene clustering method. We performed this bi-clustering method via NMF using the AD patient expression data to cluster AD patients into k sample clusters and genes into k relatively up and downregulated clusters, respectively. The T2D patient expression data were processed in the same manner.

However, in most gene expression analyses, the number of genes (features) is larger than the number of samples in the dataset. Even if the dimension of genes and samples in an expression dataset can be reduced using NMF, it is still difficult to analyze genes in k clusters because each gene belongs to one of the k clusters. In addition, each cluster may contain genes whose expression values in the samples of the given cluster are not different compared with those in samples of other clusters. Thus, some genes will be assigned to a gene cluster even if there are no relative differences in expression between sample clusters (Supplementary Fig. S1).

To address this issue, we filtered out genes in clusters by considering the original input matrix A and identified which genes in each cluster showed significantly different expression levels in a specific sample cluster. First, for each gene that was already assigned to the cluster, the distribution of expression levels was defined as $D_i (i = 1 \sim k)$ for each k sample cluster. We then used the Kruskal–Wallis test to identify genes with a significantly different expression distribution in samples of a given cluster from those in samples of at least one other cluster. The p values of the test were adjusted according to the Bonferroni correction for multiple comparisons, and genes with a q value < 0.05 were selected. Further, the Dunn test was performed between the distribution of expression levels for each gene in all possible sample cluster pairs. If the expression level differences of the gene between the given cluster and other remaining clusters were significant (q values < 0.05 after Bonferroni correction), the gene was selected. For cluster i , these genes with relatively upregulated expression were denoted as G_i^+ (Fig. 1b).

Similarly, for a gene assigned to the gene cluster by the “Min” gene clustering method, the Kruskal–Wallis test and Dunn test were subsequently applied to the genes in each cluster. For cluster i , these genes with relatively downregulated expression were selected and denoted as G_i^- . This process could effectively reduce the number of genes in each cluster compared to the conventional clustering method of NMF, which facilitated the analysis.

Differentially expressed gene (DEG) subgroups. The Kruskal–Wallis–based gene clustering method described above can extract the genes with relatively up and downregulated expression in patients with a given disease in k sample clusters. However, we further needed to identify whether genes in the obtained G_i^+ and G_i^- groups are differentially expressed in the sample cluster S_i compared to controls. In addition, even if G_i selected as the characteristic of S_i differs significantly from other sample clusters, genes in G_i need to be differentially expressed in S_i compared to controls, which means that it can also be the characteristic of the S_j sample cluster. Especially, because AD and T2D datasets are gene expression data from different tissue types, by extracting differentially expressed genes between each patient groups of AD and T2D and their healthy controls, we can remove tissue-specific genes and detect genes related to each disease. Therefore, we considered DEGs from all genes in G_i^+ and G_i^- between each patient cluster and their respective controls. First, we collected all genes assigned to any k cluster. Second, the expression levels for each gene were compared between disease samples in the i cluster and control samples using the t -test followed by Bonferroni correction. The genes with a q value $<$

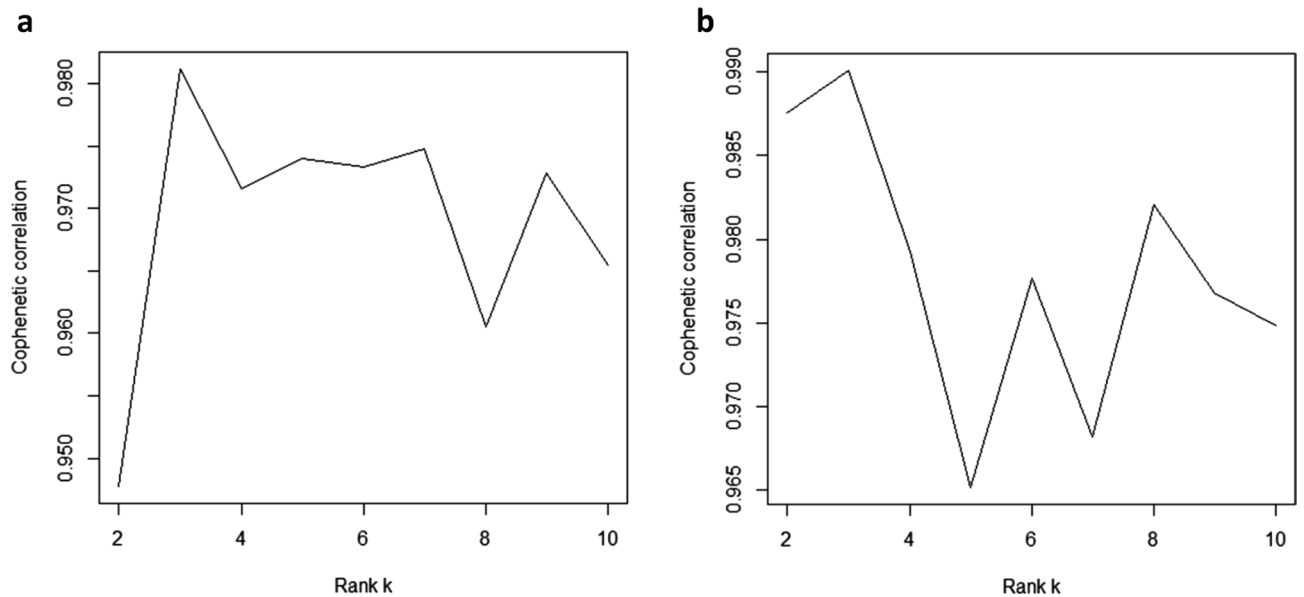


Figure 2. Cophenetic correlation coefficients for the consensus matrix in (a) Alzheimer's disease (AD) and (b) type 2 diabetes (T2D) datasets.

0.05 for the sample i cluster were assigned to M_i^+ or M_i^- depending on the direction of the expression level difference (upregulation or downregulation, respectively) (Fig. 1c).

Because AD and T2D gene expression datasets were obtained from different tissue types, this subgrouping of genes based on DEG is necessary. We can remove tissue-specific genes and select genes related to each disease by using DEGs between each patient groups of AD and T2D and their healthy controls.

Extraction of AD and T2D-related subgroup pairs and candidate genes. We independently applied the NMF approach for the expression data of AD and T2D patients and their respective controls, and obtained AD and T2D DEG subgroups for specific sample clusters. We then aimed to find AD subgroups related to T2D and T2D subgroups related to AD. To this end, we performed pathway enrichment analysis. These enriched pathways were then compared with those of known AD- and T2D-related genes in DigSee⁴¹. In addition, we investigated whether the enriched pathways in the AD DEG subgroups overlapped with T2D-related pathways and vice versa. Then, we selected a AD subgroup and a T2D subgroup containing the largest number of T2D-related pathways and AD-related pathways, respectively, which are a pair of clusters related with each other. From these two clusters, we selected common candidate genes with the same regulation direction, which are referred to as candidate genes (Fig. 1d).

Afterwards, we used an independent AD dataset downloaded from the ADNI (<http://adni.loni.usc.edu>)³¹, which included gene expression data from 116 AD patients and 246 controls extracted from peripheral blood, to validate the candidate genes identified from the DEG subgroup pairs related to both the diseases. With this dataset, the expression levels of 20,384 protein-coding genes filtered using the database Homo sapiens.GRCh38.94 were used for the classification of AD and the control sample using logistic regression. Tenfold cross-validation was performed with zero initialization and a learning rate of 0.05, and the area under the curve (AUC) was calculated at each tenfold cross-validation to evaluate the predictive ability of the candidate genes.

Additionally, we collected independent T2D gene expression datasets: 25 T2D patient and 71 control samples extracted from beta-cells or pancreatic islets in GSE20966, GSE25724, and GSE38642^{42–45}. Then, we selected protein-coding genes using Homo sapiens.GRCh38.94 from each dataset. By removing the batch effect using the “removeBatchEffect” function in the R package “limma”, we normalized the expression profiles for 10,490 common genes in the three datasets and merged them. Similar to AD, we validated the candidate genes using these T2D datasets. Because of the small number of T2D patients, we performed threefold cross-validation with zero initialization and a learning rate of 0.005 in a logistic regression model.

Results and discussion

Clustering of AD and T2D genes. The AD and T2D subgroups were independently defined using the $\log_2(\text{fold change})$ values from the expression data of 285 AD and 68 T2D patients using the convex NMF approach and NMF-based clustering method. First, to decompose the expression data of the patients into subgroups using the NMF approach, we needed to determine the optimal number of subgroups. In general, initialization of the matrices W and H affects the final outputs of NMF. Hence, we applied the NMF algorithm 10 times for each rank k from 2 to 10 with randomly initialized W and H matrices, and then calculated the average of the cophenetic correlation coefficient. We chose the rank k that had the largest average cophenetic correlation coefficient. Figure 2 shows the average cophenetic correlation coefficients for each rank k in each AD and T2D

(a) AD			
Cluster i	S_i	G_i^+	G_i^-
1	93	3004	1989
2	90	1242	1742
3	102	1129	748
(b) T2D			
Cluster i	S_i	G_i^+	G_i^-
1	17	116	144
2	9	302	564
3	42	3043	6661

Table 1. Clustered samples and genes for (a) Alzheimer’s disease (AD) and (b) type 2 diabetes (T2D).

patient dataset. The optimal rank k of both datasets was 3. Thus, we used the factorized matrices W and H with the largest cophenetic correlation coefficient out of 10 iterations of rank 3 as the NMF output for each disease.

After applying the sample and gene assignment method to the decomposed matrices, we constructed three clusters containing a subset of samples for genes with relatively upregulated and downregulated expression, respectively. For sample clustering, the “Max” method was applied to matrix H in columns, and the 285 AD patients and 68 T2D patients were divided into three sample clusters of 93, 90, and 102 patients, and 17, 9, and 42 patients, respectively. All of the genes in both datasets were first assigned to one of the three gene clusters through the “Max” method. According to the Kruskal-Wallis and Dunn test, 5375 and 4479 genes were significantly upregulated (G_x^+) and downregulated (G_y^-), respectively, in expression from other clusters for AD. Similarly, 3461 and 7369 genes were upregulated and downregulated, respectively, for T2D (Table 1). Each gene can be assigned to both G_x^+ and G_y^- when the distribution in S_x is relatively upregulated whereas that in S_y is relatively downregulated. Thus, in the union of upregulated and downregulated genes, 6729 and 10,051 genes emerged as showing significant differences in expression from other clusters for AD and T2D, respectively.

Each gene can be assigned to both G_x^+ and G_y^- when the distribution in S_x is relatively upregulated whereas that in S_y is relatively downregulated. The Kruskal-Wallis based gene clustering method showed that samples and genes in AD patients were more evenly divided compared to those in T2D patients. However, in the T2D dataset, most of the samples were assigned to S_3 and the gene cluster also showed a skewed distribution in G_3 . Because the number of clusters was decided by the optimal rank k , the cluster S_2 with the small number samples can be generated when the total number of samples are small such as the T2D dataset. This small size cluster may have distinct characteristics that can be distinguished from other clusters.

To visually confirm that the Kruskal-Wallis-based gene clustering method removed inappropriate genes in each gene cluster compared to the conventional “Max”/“Min” method, the gene expression matrices were rearranged in cluster order for both genes and samples, which were visualized on a heatmap. The rectangular regions on the diagonal of the heatmap, indicating samples and genes assigned in the same cluster, demonstrate genes with relatively upregulated or downregulated expression in each cluster. Compared to the basic “Max” and “Min” method, genes selected by Kruskal-Wallis test generated more distinct regions, in which genes showed significant differences in expression levels that could be clearly observed on the heatmap for both the AD and T2D datasets (Fig. 3). Figure 3 was generated using “heatmap” function in the R package “NMF”⁴⁶ and “heatmap” function in the python library “seaborn”⁴⁷.

Next, we constructed M_i^+ and M_i^- , where gene expressions differ between each disease patient group and its corresponding control group. For the subgroups of each sample cluster, we integrated the genes with relatively up and downregulated expression compared to patients in other clusters (G_i^+ and G_i^-), resulting in 6729 and 10,051 integrated genes in AD and T2D, respectively. Then, genes were assigned into a subgroup i if they are differentially expressed between patients in the subgroup i and control samples based on the t -test with Bonferroni-corrected q value < 0.05 . The numbers of these up and downregulated genes (M_i^+ and M_i^-) are shown in Table 2.

We further examined the clinical characteristics such as age of these samples in AD modules because the age information is only in AD, not T2D. The result of a one-way ANOVA for age in the AD patient subgroup showed no significant differences between AD patient subgroups (Supplementary Table S2), which suggests that the grouping of samples does not depend on age.

Selecting cluster pairs of interest. To find AD clusters related to T2D and T2D clusters related to AD, pathway enrichment analysis was performed for the up and downregulated genes (M_i^+ and M_i^-) in each AD and T2D cluster using a total of 10,378 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) terms from MSigDB. The hypergeometric test for M_i^+ and M_i^- in AD and T2D genes was used to determine significantly enriched KEGG pathways and GO terms (Tables 3 and 4). As references of AD- and T2D-related pathways, we extracted 1635 AD-related genes and 1658 T2D-related genes from DigSee⁴¹, and obtained 1675 and 1757 AD- and T2D-related pathways, respectively, using the hypergeometric enrichment test. We identified common pathways from the DEG subgroups in AD patients with 1757 T2D-related pathways from DigSee⁴¹. Among the AD clusters, patients in S_3 were most likely to have an association with T2D compared to AD patients in the AD S_1 and AD S_2 clusters (Table 3). Likewise, T2D patients in cluster S_3 were most likely to be associated with AD (Table 4). There was no clinical information in the datasets to confirm whether the AD S_3

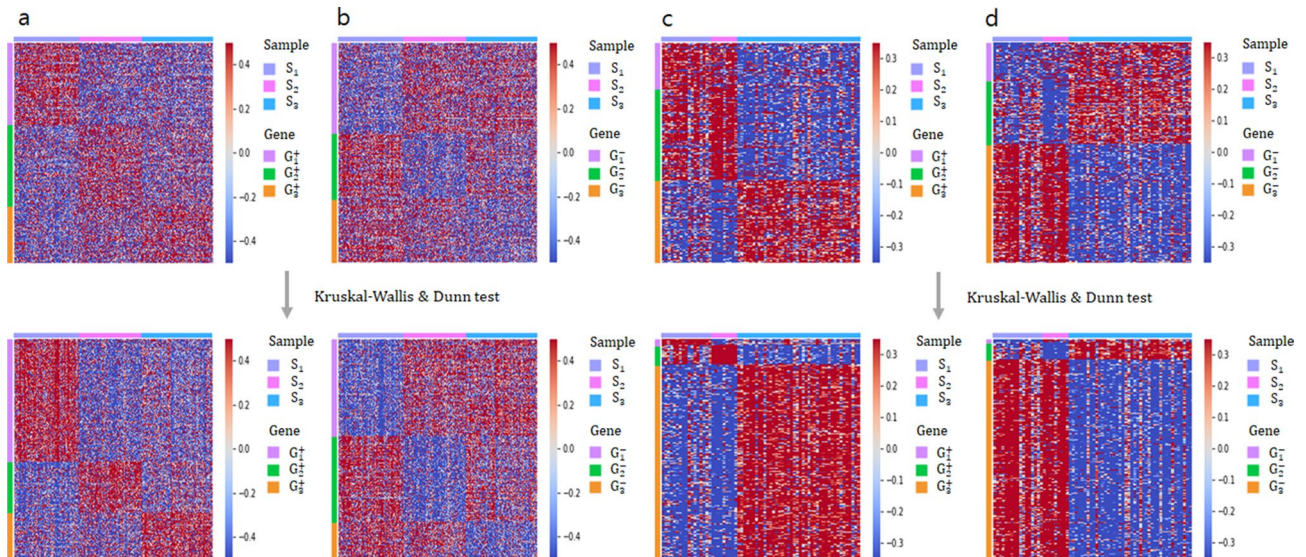


Figure 3. Heatmap of gene clustering method in Alzheimer’s disease (AD) and type 2 diabetes (T2D). (a) and (b) show genes and sample clusters from AD patients. (c) and (d) show genes and sample clusters from T2D patients. Top panels show clusters obtained using the basic “Max” ((a) and (c)) and “Min” ((b) and (d)) methods. Bottom panels show clusters obtained using the Kruskal–Wallis and Denn test for relatively upregulated genes ((a) and (c)) and relatively downregulated genes ((b) and (d)). This figure is generated by the R software (R version 3.6.1, <https://www.r-project.org/>) and python (version 3.6.8, <https://www.python.org/>).

(a) AD			
Cluster i	S_i	M_i^+	M_i^-
1	93	1281	825
2	90	1443	2422
3	102	1482	1670
(b) T2D			
Cluster i	S_i	M_i^+	M_i^-
1	17	228	426
2	9	41	283
3	42	1235	4113

Table 2. Number of DEGs in each cluster.

AD M_i	Enriched pathways	Common pathways with 1757 T2D-related pathways
M_1^+	16	0
M_1^-	10	5
M_2^+	30	9
M_2^-	101	4
M_3^+	205	119
M_3^-	164	8

Table 3. Numbers of type 2 diabetes (T2D)-related pathways for each Alzheimer’s disease (AD) differentially expressed gene module.

patients actually have T2D or whether the T2D S_3 patients have AD; however, our data suggest that patients in these clusters might share genetic characteristics of the other disease.

In addition, we used brain gene expression data GSE5281 to determine which AD clusters have features in common with the AD brain⁴⁸. We extracted 1831 DEGs from these gene expression data by performing the t -test with a q value < 0.05, and 160 pathways were extracted from these genes. As a result, genes in the AD S_3 module most overlapped with AD brain-related pathways (Supplementary Table S1).

To obtain further evidence that AD S_3 and T2D S_3 are related, pathway enrichment analysis was performed for the common genes of possible cluster pairs in the up-regulated and down-regulated AD and T2D gene clusters.

T2D M_i	Enriched pathways	Common pathways with 1675 AD-related pathways
M_1^+	4	0
M_1^-	5	5
M_2^+	0	0
M_2^-	0	0
M_3^+	135	31
M_3^-	87	16

Table 4. Numbers of Alzheimer's disease (AD)-related pathways in each type 2 diabetes (T2D) differentially expressed gene module.

AD M_i	T2D M_j	Pathways from common genes	Common pathways with DigSee
M_1^+	M_1^+	0	0
M_1^-	M_1^-	2	0
M_1^+	M_2^+	0	0
M_1^-	M_2^-	0	0
M_1^+	M_3^+	0	0
M_1^-	M_3^-	0	0
M_2^+	M_1^+	0	0
M_2^-	M_1^-	0	0
M_2^+	M_2^+	0	0
M_2^-	M_2^-	0	0
M_2^+	M_3^+	23	3
M_2^-	M_3^-	0	0
M_3^+	M_1^+	0	0
M_3^-	M_1^-	0	0
M_3^+	M_2^+	0	0
M_3^-	M_2^-	1	0
M_3^+	M_3^+	21	12
M_3^-	M_3^-	18	7

Table 5. Numbers of common pathways between disease subgroup pairs and DigSee.

The common genes with the same regulation direction (i.e., up or downregulated) in both diseases were extracted, and their enriched pathways were compared with the 1498 pathways of the 671 intersecting genes between AD and T2D from DigSee⁴¹. We used the pathways associated with both AD and T2D as references to identify whether the common genes in each possible cluster pair were related. Indeed, the enriched pathways from common genes in the upregulated AD M_3^+ and T2D M_3^+ modules showed more overlap with pathways from DigSee compared to that in other possible pairs (Table 5). Similarly, the common genes from the downregulated AD M_3^- and T2D M_3^- modules showed the highest overlapping pathway ratio with pathways from DigSee (Table 5). This suggested that the cluster pairs AD M_3 and T2D M_3 were the most closely related to both AD and T2D.

Extraction of candidate genes. We extracted 241 common genes from the cluster pair AD M_3 and T2D M_3 , including 195 upregulated genes and 46 downregulated genes, which were selected as candidate genes associated with both AD and T2D (Supplementary Table S3). Among these 241 genes, 14 genes were common with genes related with both AD and T2D from DigSee⁴¹. In DigSee, 661 genes were common for both AD and T2D. With a hypergeometric test for significance of 14 genes out of 241, a p value was 0.03826, showing significance of these genes in their roles in AD and T2D. In addition, we collected more AD and T2D genes from AlzGene and T2DiACoD (Supplementary Table S3)^{49,50}. When comparing the 241 genes with genes in the three databases of DigSee, AlzGene, and T2DiACoD, 56 genes were related to AD or T2D.

These candidate genes were enriched in 29 pathways (Supplementary Table S4) and 14 pathways that are common with AD and T2D-related pathways from DigSee are shown in Table 6. Pathways associated with common pathological features of AD and T2D such as the immune system-related pathways (T cell selection, positive T cell selection, and T cell differentiation)^{51,52} and chemokine signaling pathway^{53,54} were included. Immune system-related pathways are known to be common characteristics of AD in the brain and blood⁵¹, and there are some evidences that chemokines play an essential role in the central nervous system and neuroprotection^{55,56}. Interestingly, 11 genes among the 241 candidate genes were involved in chemokine signaling pathway, and 6 of them were AD and T2D-related genes: RAF1, RAC1, RHOA, STAT3, AKT1, and PRKCD.

Pathways	<i>P</i> value	Adjusted <i>p</i> value
GO: Positive regulation of cell-cell adhesion	3.79E-09	2.85E-05
KEGG: Pancreatic cancer	2.17E-07	4.03E-05
GO: T cell selection	5.83E-09	4.39E-05
GO: Positive T cell selection	7.33E-09	5.52E-05
GO: T cell differentiation	5.64E-08	0.000425056
KEGG: Adherens junction	3.57E-06	0.000663158
KEGG: Epithelial cell signaling in Helicobacter pylori infection	2.23E-05	0.004142846
KEGG: Chemokine signaling pathway	2.58E-05	0.004806622
KEGG: JAK-STAT signaling pathway	2.62E-05	0.004876146
GO: Vacuolar lumen	1.12E-05	0.011236912
KEGG: T cell receptor signaling pathway	6.41E-05	0.011931807
KEGG: Colorectal cancer	0.000123344	0.022942068
KEGG: Neurotrophin signaling pathway	0.000188801	0.035116995
KEGG: Fc gamma R-mediated phagocytosis	0.00020364	0.037877102

Table 6. Common pathways between candidate genes and DigSee genes.

(AD M_i , T2D M_j)	AUC
(M_1, M_1)	0.5173
(M_1, M_2)	0.6034
(M_1, M_3)	0.5810
(M_2, M_1)	0.5264
(M_2, M_2)	0.5763
(M_2, M_3)	0.6256
(M_3, M_1)	0.5411
(M_3, M_2)	0.6022
(M_3, M_3)	0.6906

Table 7. Performance of classification of Alzheimer's disease (AD) and controls in the ADNI cohort using different sets of genes. (M_i, M_j) represents the common genes between the AD M_i module and type 2 diabetes (T2D) M_j module used for classification.

To verify whether the 241 candidate genes could be informative markers for the classification of each disease patients and controls, we used data from the ADNI cohort³¹ for AD prediction, and a merged independent T2D dataset for predicting T2D. In AD prediction, candidate genes from the (AD M_3 , T2D M_3) pair showed the best diagnostic performance, with an AUC value of 0.6906, compared with genes from nine possible pairs (Table 7).

For comparison, the performance of classifying AD in the ADNI cohort was measured for 250 random genes to match the size of the candidate genes. Classification using 250 random genes was performed 100 times, and the mean AUC value was 0.5658. The *t*-test showed that the candidate genes significantly outperformed the randomly selected genes in classification with a *p* value of 5.723×10^{-52} (Fig. 4). As another comparison, we obtained 1,466 DEGs from the AD samples (GSE63060 and GSE63061 datasets) by the *t*-test and selected genes with a *q* value < 0.05 after Bonferroni correction. When these DEGs were used for AD classification in the ADNI cohort, the AUC value was 0.5757.

Furthermore, we examined serum glucose data in the ADNI dataset to consider the clinical characteristics of samples. When we compared glucose levels between the AD and the control groups by the *t*-test, there was no significant difference (*p* value = 0.513). In ADNI, there were 41 AD patients and 95 controls with high glucose levels, including prediabetes or diabetes samples (≥ 100 mg/dL). Using the logistic regression model that were constructed for classification of AD (Table 7), we examined the classification performance for these hyperglycemic samples in the test set of each fold on tenfold cross-validation. The prediction performance for these hyperglycemic samples using the candidate genes (AUC = 0.715) was the highest among those using other genes (Supplementary Table S5). We also found that the predictive power of the prediction model using candidate genes was higher for these hyperglycemic samples compared to those for the whole samples (0.715 in Supplementary Table S5 and 0.6906 in Table 7, respectively).

We also performed classification using the independent T2D datasets (25 T2D and 71 controls). Among the 241 candidate genes, 179 genes were included in the T2D samples. On the threefold cross-validation, we obtained the AUC value of 0.9543. The AUC value of randomly selected 180 genes was 0.9458 and the predictive performances using other possible gene pairs were also high (Supplementary Table S6). This indicates that gene expression levels between T2D samples and controls in pancreatic islets were significantly different for most genes.

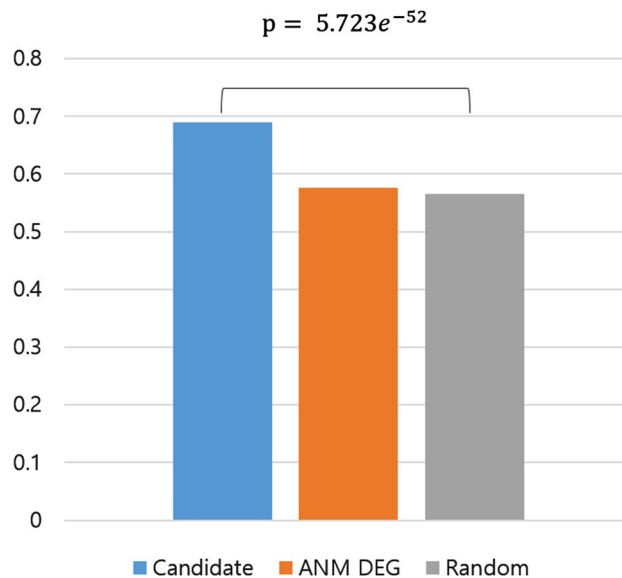


Figure 4. Performance comparison of Alzheimer's disease (AD) classification from controls based on the area under the curve (AUC) using different gene sets.

Application of the proposed model to another AD dataset. We applied the proposed model to another AD dataset. For AD, ADNI was used instead of gene expression datasets of GSE63060 and GSE63061. For T2D, the same gene expression data (GSE78721) was used. When we try to find the optimal number of clusters for the ADNI dataset, the optimal rank k of the ADNI dataset was 2 (Supplementary Fig. S2). At least three clusters are required to determine significant differences in the distribution of each gene between clusters with the Kruskal-Wallis and Dunn test. Thus, as an alternative, we clustered the ADNI data with the non-optimal rank $k = 3$, resulting that 116 AD patients were clustered into three clusters with 17, 54, and 45 samples, respectively (Supplementary Table S7). When the pathway enrichment analysis was performed for genes in each cluster, the ADNI M_3^+ gene cluster contained the largest number of T2D-related pathways and followed by M_1^+ . Among 41 hyperglycemic AD patients, the largest 19 belonged to ADNI S3. However, the proportion of hyperglycemic AD patients in each ADNI sample cluster was the highest in ADNI S1 followed by S3 ($S1 = 0.47\%$, $S2 = 0.259\%$, and $S3 = 0.422\%$), which implies that the characteristics of patients of S_1 and S_3 are similar and can be merged for the high risk subgroup of T2D.

Additionally, there was no difference between the age of patients in S_1 and S_3 , but the age of patients of S_2 was significantly lower than those of S_1 and S_3 (p values were 0.0013 and 0.0207 with a one-way ANOVA test, respectively). We also performed a one-way ANOVA test for APOE4 among subgroups of AD patients and observed no significant difference between APOE4 (p value as 0.35). Therefore, the proposed method may cluster patients that have some similar clinical characteristics such as the age, but not all of subgroups were clustered by these characteristics.

Conclusion

We have provided a methodological and analytical approach for identifying correlations between AD and T2D at the genetic level. Since the AD dataset does not contain information about whether the AD patients have T2D or not, it is important to define subgroups of AD; the same is true for T2D. Because the conventional NMF is not suitable for this task, we developed a method of gene selection from gene expression data. After applying NMF to gene expression data, additional conditions were taken into account for detecting distinct characteristics of subgroups. Genes with significant differences in expression levels in each patient groups (AD and T2D) were first selected to screen patients with AD associated with T2D and patients with T2D associated with AD. We identified genes that characterize these specific AD and T2D patients and identified the potential relationship between the two diseases based on gene expression profiles. To validate these potential relationships from candidate genes, prediction errors of the classification between AD and controls from logistic regression were compared with randomly selected genes in an independent AD dataset. Inclusion of the candidate genes significantly increased the AUC values in classifying AD from controls compared with randomly selected genes.

In conclusion, we provide new insights for extracting differentially expressed genes with relative differences in a specific patient group. These genes were enriched with pathways related to both AD and T2D such as T cell selection and chemokine pathways. As AD patients have genetic heterogeneity, the investigation of commonly dysregulated pathways in AD and T2D can enhance personalized medical cares for a subgroup of AD. Further studies are needed to reveal the relationships among AD and other AD-related diseases which could improve the prevention and treatment of AD.

Received: 20 January 2021; Accepted: 24 June 2021

Published online: 27 July 2021

References

- Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement.* **3**, 186–191 (2007).
- Barker, W. W. *et al.* Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the state of Florida Brain Bank. *Alzheimer Dis. Assoc. Disord.* **16**, 203–212 (2002).
- van Oijen, M., de Jong, F. J., Hofman, A., Koudstaal, P. J. & Breteler, M. M. Subjective memory complaints, education, and risk of Alzheimer's disease. *Alzheimer's Dement.* **3**, 92–97 (2007).
- Armstrong, R. A. The molecular biology of senile plaques and neurofibrillary tangles in Alzheimer's disease. *Folia Neuropathol.* **47**, 289–99 (2009).
- Gouras, G. K., Olsson, T. T. & Hansson, O. β -amyloid peptides and amyloid plaques in Alzheimer's disease. *Neurotherapeutics* **12**, 3–11 (2015).
- Vogel, G. Tau protein mutations confirmed as neuron killers. *Science* **280**, 1524–1525 (1998).
- Grant, W. B., Campbell, A., Itzhaki, R. F. & Savory, J. The significance of environmental factors in the etiology of Alzheimer's disease. *J. Alzheimers Dis.* **4**, 179–189 (2002).
- Abid, N. B., Naseer, M. I. & Kim, M. O. Comparative gene-expression analysis of Alzheimer's disease progression with aging in transgenic mouse model. *Int. J. Mol. Sci.* **20**, 1219 (2019).
- Sherva, R. *et al.* Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimer's Dement.* **10**, 45–52 (2014).
- Mayeux, R. & Schupf, N. Blood-based biomarkers for Alzheimer's disease: Plasma $\alpha\beta 40$ and $\alpha\beta 42$, and genetic variants. *Neurobiol. Aging* **32**, S10–S19 (2011).
- Loring, J., Wen, X., Lee, J., Seilhamer, J. & Somogyi, R. A gene expression profile of Alzheimer's disease. *DNA Cell Biol.* **20**, 683–695 (2001).
- Khan, A. T., Dobson, R. J., Sattlecker, M. & Kiddle, S. J. Alzheimer's disease: Are blood and brain markers related? A systematic review. *Ann. Clin. Transl. Neurol.* **3**, 455–462 (2016).
- Bu, X. *et al.* Blood-derived amyloid- β protein induces Alzheimer's disease pathologies. *Mol. Psychiatry* **23**, 1948–1956 (2018).
- Fehlbaum-Beurdeley, P. *et al.* Toward an Alzheimer's disease diagnosis via high-resolution blood gene expression. *Alzheimer's Dement.* **6**, 25–38 (2010).
- Booij, B. B. *et al.* A gene expression pattern in blood for the early detection of Alzheimer's disease. *J. Alzheimers Dis.* **23**, 109–119 (2011).
- Lunnon, K. *et al.* A blood gene expression marker of early Alzheimer's disease. *J. Alzheimers Dis.* **33**, 737–753 (2013).
- Rye, P. *et al.* A novel blood test for the early detection of Alzheimer's disease. *J. Alzheimers Dis.* **23**, 121–129 (2011).
- Fratiglioni, L. *et al.* Prevalence of Alzheimer's disease and other dementias in an elderly urban population: Relationship with age, sex, and education. *Neurology* **41**, 1886–1886 (1991).
- Nussbaum, R. L. & Ellis, C. E. Alzheimer's disease and Parkinson's disease. *N. Engl. J. Med.* **348**, 1356–1364 (2003).
- Huang, C.-C. *et al.* Diabetes mellitus and the risk of Alzheimer's disease: A nationwide population-based study. *PLoS One* **9**, e87095 (2014).
- Panigrahi, P. P. & Singh, T. R. Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: Revealed association with aging and other diseases. *J. Theor. Biol.* **334**, 109–121 (2013).
- Edwards, G. A. III., Gamez, N., Escobedo, G. Jr., Calderon, O. & Moreno-Gonzalez, I. Modifiable risk factors for Alzheimer's disease. *Front. Aging Neurosci.* **11**, 146 (2019).
- Santiago, J. A. & Potashkin, J. A. The impact of disease comorbidities in Alzheimer's disease. *Front. Aging Neurosci.* **13**, 38 (2021).
- Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Mol. Syst. Biol.* **10**, 743 (2014).
- Barbagallo, M. & Dominguez, L. J. Type 2 diabetes mellitus and Alzheimer's disease. *World J. Diabetes* **5**, 889 (2014).
- Haataja, L., Gurlo, T., Huang, C. J. & Butler, P. C. Islet amyloid in type 2 diabetes, and the toxic oligomer hypothesis. *Endocr. Rev.* **29**, 303–316 (2008).
- Benilova, I., Karran, E. & De Strooper, B. The toxic $\alpha\beta$ oligomer and Alzheimer's disease: An emperor in need of clothes. *Nat. Neurosci.* **15**, 349–357 (2012).
- De Felice, F. G. & Ferreira, S. T. Inflammation, defective insulin signaling, and mitochondrial dysfunction as common molecular denominators connecting type 2 diabetes to Alzheimer disease. *Diabetes* **63**, 2262–2272 (2014).
- De Felice, F. G., Lourenco, M. V. & Ferreira, S. T. How does brain insulin resistance develop in Alzheimer's disease?. *Alzheimer's Dement.* **10**, S26–S32 (2014).
- Götz, J., Ittner, L. & Lim, Y.-A. Common features between diabetes mellitus and Alzheimer's disease. *Cell. Mol. Life Sci.* **66**, 1321–1325 (2009).
- Mueller, S. G. *et al.* Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI). *Alzheimer's Dement.* **1**, 55–66 (2005).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).
- Lee, M.-W., Lee, M. & Oh, K.-J. Adipose tissue-derived signatures for obesity and type 2 diabetes: Adipokines, batokines and microRNAs. *J. Clin. Med.* **8**, 854 (2019).
- Tiwari, P. *et al.* Systems genomics of thigh adipose tissue from Asian Indian type-2 diabetics revealed distinct protein interaction hubs. *Front. Genet.* **9**, 679 (2019).
- Chiaromonte, F. & Martinelli, J. Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.* **176**, 123–144 (2002).
- Liu, W., Yuan, K. & Ye, D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *J. Biomed. Inform.* **41**, 602–606 (2008).
- Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**, 4164–4169 (2004).
- Carmona-Saez, P., Pascual-Marqui, R. D., Tirado, F., Carazo, J. M. & Pascual-Montano, A. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinform.* **7**, 78 (2006).
- Mejía-Roa, E. *et al.* Biclustering and classification analysis in gene expression using nonnegative matrix factorization on multi-GPU systems. In *2011 11th International Conference on Intelligent Systems Design and Applications*, 882–887 (IEEE, 2011).
- Ding, C. H., Li, T. & Jordan, M. I. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 45–55 (2008).
- Kim, J. *et al.* Digsee: Disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Res.* **41**, W510–W517 (2013).

42. Marselli, L. *et al.* Gene expression profiles of beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. *PLoS One* **5**, e11499 (2010).
43. Dominguez, V. *et al.* Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic β cells. *J. Biol. Chem.* **286**, 4216–4225 (2011).
44. Taneera, J. *et al.* A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab.* **16**, 122–134 (2012).
45. Zhong, M., Wu, Y., Ou, W., Huang, L. & Yang, L. Identification of key genes involved in type 2 diabetic islet dysfunction: A bioinformatics study. *Biosci. Rep.* **39**, BSR20182172 (2019).
46. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **11**, 1–9 (2010).
47. Waskom, M. *et al.* Mwashom/seaborn: V0. 8.1 (september 2017). *Zenodo* (2017).
48. Liang, W. S. *et al.* Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol. Genomics* **28**, 311–322 (2007).
49. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).
50. Rani, J. *et al.* T2diacod: A gene atlas of type 2 diabetes mellitus associated complex disorders. *Sci. Rep.* **7**, 1–21 (2017).
51. Jevtic, S., Sengar, A. S., Salter, M. W. & McLaurin, J. The role of the immune system in Alzheimer disease: Etiology and treatment. *Ageing Res. Rev.* **40**, 84–94 (2017).
52. Berbudi, A., Rahmadika, N., Tjahjadi, A. I. & Ruslami, R. Type 2 diabetes and its impact on the immune system. *Curr. Diabetes Rev.* **16**, 442 (2020).
53. Zuena, A. R., Casolini, P., Lattanzi, R. & Maftei, D. Chemokines in Alzheimer's disease: New insights into prokineticins, chemokine-like proteins. *Front. Pharmacol.* **10**, 622 (2019).
54. Yao, L., Herlea-Pana, O., Heuser-Baker, J., Chen, Y. & Barlic-Dicen, J. Roles of the chemokine system in development of obesity, insulin resistance, and cardiovascular disease. *J. Immunol. Res.* **2014**, 181450 (2014).
55. Bajetto, A., Bonavia, R., Barbero, S., Florio, T. & Schettini, G. Chemokines and their receptors in the central nervous system. *Front. Neuroendocrinol.* **22**, 147–184 (2001).
56. Jaerve, A. & Müller, H. W. Chemokines in CNS injury and repair. *Cell Tissue Res.* **349**, 229–248 (2012).

Acknowledgements

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found http://adni.loni.usc.edu/wp-content/themes/freshnewsdev-v2/documents/policy/ADNI_Acknowledgement_List%205-29-18.pdf. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Samples from the National Cell Repository for AD (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (AIG), were used in this study. Additional support for data analysis was provided by NLM R01 LM012535, NIA R03 AG054936, and the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Author contributions

H.L. contributed to the study concept and design. Y.C. designed and implemented the proposed algorithms. H.L. and Y.C. analyzed and interpreted results. Y.C. and H.L. wrote the manuscript. H.L. took part in the study supervision and coordination. All authors read and approved the final manuscript.

Funding

This work has been supported by the Bio & Medical Technology Development Program of NRF funded by the Korean government (MSIT) (NRF-2018M3C7A1054935) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant Number: HI18C0460).

Competing interests

The authors declare no competing interests

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-94048-0>.

Correspondence and requests for materials should be addressed to H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

the Alzheimer's Disease Neuroimaging Initiative

Michael W. Weiner², Paul Aisen³, Ronald Petersen⁴, Clifford R. Jack Jr.⁴, William Jagust⁵, John Q. Trojanowki⁶, Arthur W. Toga⁷, Laurel Beckett⁸, Robert C. Green⁹, Andrew J. Saykin¹⁰, John Morris¹¹, Leslie M. Shaw⁶, Zaven Khachaturian¹², Greg Sorensen¹³, Maria Carrillo¹⁴, Lew Kuller¹⁵, Marc Raichle¹¹, Steven Paul¹⁶, Peter Davies¹⁷, Howard Fillit¹⁸, Franz Hefti¹⁹, Davie Holtzman¹¹, M. Marcel Mesulam²⁰, William Potter²¹, Peter Snyder²², Tom Montine²³, Ronald G. Thomas³, Michael Donohue³, Sarah Walter³, Tamie Sather³, Gus Jiminez³, Archana B. Balasubramanian³, Jennifer Mason³, Iris Sim³, Danielle Harvey⁸, Matthew Bernstein⁴, Nick Fox²⁴, Paul Thompson²⁵, Norbert Schuff², Charles DeCarli⁸, Bret Borowski⁴, Jef Gunter⁴, Matt Senjem⁴, Prashanthi Vemuri⁴, David Jones⁴, Kejal Kantarci⁴, Chad Ward⁴, Robert A. Koeppe²⁶, Norm Foster²⁷, Eric M. Reiman²⁸, Kewei Chen²⁸, Chet Mathis¹⁵, Susan Landau⁵, Nigel J. Cairns¹¹, Erin Householder¹¹, Lisa Taylor-Reinwald¹¹, Virginia Lee²⁵, Magdalena Korecka²⁵, Michal Figurski²⁵, Karen Crawford⁷, Scott Neu⁷, Tatiana M. Foroud¹⁰, Steven Potkin²⁹, Li Shen¹⁰, Kelley Faber¹⁰, Sungeun Kim¹⁰, Lean Tha¹³, Richard Frank³¹, John Hsiao³², Jeffrey Kaye³³, Joseph Quinn³⁴, Lisa Silbert³⁴, Betty Lind³⁴, Raina Carter³⁴, Sara Dolen³⁴, Beau Ances¹¹, Maria Carroll¹¹, Mary L. Creech¹¹, Erin Franklin¹¹, Mark A. Mintun¹¹, Stacy Schneider¹¹, Angela Oliver¹¹, Lon S. Schneider¹¹, Sonia Pawluczyk⁷, Mauricio Beccera⁷, Liberty Teodoro⁷, Bryan M. Spann⁷, James Brewer³, Helen Vanderswag³, Adam Fleisher³, Daniel Marson³⁴, Randall Griffith³⁴, David Clark³⁴, David Geldmacher³⁴, John Brockington³⁴, Erik Roberson³⁴, Marissa Natelson Love³⁰, Judith L. Heidebrink⁵, Joanne L. Lord⁵, Sara S. Mason⁴, Colleen S. Albers⁴, David Knopman⁴, Kris Johnson⁴, Hillel Grossman³¹, Efe Mitsis³⁵, Raj C. Shah³⁶, Leyla deToledo-Morrell³⁶, Rachelle S. Doody³⁷, Javier Villanueva-Meyer³⁷, Munir Chowdhury³⁷, Susan Rountree³⁷, Mimi Dang³⁷, Ranjan Duara³⁸, Daniel Varon³⁸, Maria T. Greig³⁸, Peggy Roberts³⁸, Yaakov Stern³⁹, Lawrence S. Honig³⁹, Karen L. Bell³⁹, Marilyn Albert⁴⁰, Chiadi Onyike³⁰, Daniel D'Agostino II³⁰, Stephanie Kielb³⁰, James E. Galvin⁴⁰, Brittany Cerbone⁴⁰, Christina A. Michel⁴⁰, Dana M. Pogorelec⁴⁰, Henry Rusinek⁴⁰, Mony J. de Leon⁴⁰, Lidia Glodzik⁴⁰, Susan De Santi⁴⁰, Kyle Womack⁴¹, Dana Mathews⁴¹, Mary Quiceno⁴¹, P. Murali Doraiswamy⁴², Jeffrey R. Petrella⁴², Salvador Borges-Neto⁴², Terence Z. Wong⁴², Edward Coleman⁴², Allan I. Levey⁴³, James J. Lah⁴³, Janet S. Cella⁴³, Jeffrey M. Burns⁴⁴, Russell H. Swerdlow⁴⁴, William M. Brooks⁴⁴, Steven E. Arnold⁶, Jason H. Karlawish⁶, David Wolk⁶, Christopher M. Clark⁶, Liana Apostolova²⁵, Kathleen Tingus²⁵, Ellen Woo²⁵, Daniel H. S. Silverman²⁵, Po H. Lu²⁵, George Bartzokis²⁵, Charles D. Smith⁴⁵, Greg Jicha⁴⁵, Peter Hardy⁴⁵, Partha Sinha⁴⁵, Elizabeth Oates⁴⁵, Gary Conrad⁴⁵, Neill R. Graf-Radford⁴⁶, Francine Parfitt⁴⁶, Tracy Kendall⁴⁶, Heather Johnson⁴⁶, Oscar L. Lopez¹⁵, MaryAnn Oakley¹⁵, Donna M. Simpson¹⁵, Martin R. Farlow¹⁰, Ann Marie Hake¹⁰, Brandy R. Matthews¹⁰, Jared R. Brosch¹⁰, Scott Herring¹⁰, Cynthia Hunt¹⁰, Anton P. Porsteinsson⁴⁷, Bonnie S. Goldstein⁴⁷, Kim Martin⁴⁷, Kelly M. Makino⁴⁷, M. Saleem Ismail⁴⁷, Connie Brand⁴⁷, Ruth A. Mulnard⁴⁷, Gaby Thai⁴⁷, Catherine Mc-Adams-Ortiz⁴⁷, Christopher H. van Dyck⁴⁸, Richard E. Carson⁴⁸, Martha G. MacAvoy⁴⁸, Pradeep Varma⁴⁸, Howard Chertkow⁴⁹, Howard Bergman⁴⁹, Chris Hosein⁴⁹, Sandra Black⁵⁰, Bojana Stefanovic⁵⁰, Curtis Caldwell⁵⁰, Ging-Yuek Robin Hsiung⁵¹, Howard Feldman⁵¹, Benita Mudge⁵¹, Michele Assaly⁵¹, Elizabeth Finger⁵², Stephen Pasternack⁵², Irina Rachisky⁵², Dick Trost⁵², Andrew Kertesz^{52,61},

Charles Bernick⁵³, Donna Munic⁵³, Kristine Lipowski²⁰, Masandra Weintraub²⁰, Borna Bonakdarpour²⁰, Diana Kerwin²⁰, Chuang-Kuo Wu²⁰, Nancy Johnson²⁰, Carl Sadowsky⁵⁴, Teresa Villena⁵⁴, Raymond Scott Turner⁵⁵, Kathleen Johnson⁵⁵, Brigid Reynolds⁵⁵, Reisa A. Sperling⁹, Keith A. Johnson⁹, Gad Marshall⁹, Jerome Yesavage⁵⁶, Joy L. Taylor⁵⁶, Barton Lane⁵⁶, Allyson Rosen⁵⁶, Jared Tinklenberg⁵⁶, Marwan N. Sabbagh²⁸, Christine M. Belden²⁸, Sandra A. Jacobson²⁸, Sherye A. Sirrel²⁸, Neil Kowall⁵⁷, Ronald Killiany⁵⁷, Andrew E. Budson⁵⁷, Alexander Norbash⁵⁷, Patricia Lynn Johnson⁵⁷, Thomas O. Obisesan⁵⁸, Saba Wolday⁵⁸, Joanne Allard⁵⁸, Alan Lerner⁵⁹, Paula Ogrocki⁵⁹, Curtis Tatsuoka⁵⁹, Parianne Fatica⁵⁹, Evan Fletcher⁸, Pauline Maillard⁸, John Olichney⁸, Owen Carmichael⁸, Smita Kittur⁶⁰, Michael Borrie⁶¹, T.-Y. Lee⁶¹, Rob Bartha⁶¹, Sterling Johnson⁶¹, Sanjay Asthana⁶¹, Cynthia M. Carlsson⁶¹, Adrian Preda²⁹, Dana Nguyen²⁹, Pierre Tariot²⁸, Anna Burke²⁸, Nadira Trncic²⁸, Adam Fleisher²⁸, Stephanie Reeder²⁸, Vernice Bates⁶², Horacio Capote⁶², Michelle Rainka⁶², Douglas W. Scharre⁶³, Maria Katakaki⁶³, Anahita Adeli⁶³, Earl A. Zimmerman⁶⁴, Dzintra Celmins⁶⁴, Alice D. Brown⁶⁴, Godfrey D. Pearlson⁶⁵, Karen Blank⁶⁵, Karen Anderson⁶⁵, Laura A. Flashman⁶⁶, Marc Seltzer⁶⁶, Mary L. Hynes⁶⁶, Robert B. Santulli⁶⁶, Kaycee M. Sink⁶⁷, Leslie Gordineer⁶⁷, Jef D. Williamson⁶⁷, Pradeep Garg⁶⁷, Franklin Watkins⁶⁷, Brian R. Ott²², Henry Querfurth²², Geoffrey Tremont²², Stephen Salloway²², Paul Malloy²², Stephen Correia²², Howard J. Rosen², Bruce L. Miller², David Perry², Jacobo Mintzer⁶⁸, Kenneth Spicer⁶⁸, David Bachman⁶⁸, Elizabeth Finger⁶¹, Stephen Pasternak⁶¹, Irina Rachinsky⁶¹, John Rogers⁶¹, Dick Drost⁶¹, Nunzio Pomara⁶⁹, Raymundo Hernando⁶⁹, Antero Sarrael⁶⁹, Susan K. Schultz⁷⁰, Laura L. Boles Ponto⁷⁰, Hyungsub Shim⁷⁰, Karen Ekstam Smith⁷⁰, Norman Relkin¹⁶, Gloria Chaing¹⁶, Michael Lin¹⁶, Lisa Ravdin¹⁶, Amanda Smith⁷¹, Balebail Ashok Raj⁷¹ & Kristin Fargher⁷¹

²UC San Francisco, San Francisco, CA 94107, USA. ³UC San Diego, La Jolla, CA 92093, USA. ⁴Mayo Clinic, Rochester, MN, USA. ⁵UC Berkeley, Berkeley, San Francisco, USA. ⁶University of Pennsylvania, Philadelphia, PA 19104, USA. ⁷USC, Los Angeles, CA 90032, USA. ⁸UC Davis, Sacramento, CA, USA. ⁹Brigham and Women's Hospital/Harvard Medical School, Boston, MA 02215, USA. ¹⁰Indiana University, Bloomington, IN 47405, USA. ¹¹Washington University St. Louis, St. Louis, MO 63110, USA. ¹²Prevent Alzheimer's Disease, Rockville, MD 20850, USA. ¹³Siemens, Erlangen, Germany. ¹⁴Alzheimer's Association, Chicago, IL 60631, USA. ¹⁵University of Pittsburgh, Pittsburgh, PA 15213, USA. ¹⁶Cornell University, Ithaca, NY 14853, USA. ¹⁷Albert Einstein College of Medicine, Yeshiva University, Bronx, NY 10461, USA. ¹⁸AD Drug Discovery Foundation, New York, NY 10019, USA. ¹⁹Acumen Pharmaceuticals, Livermore, CA 94551, USA. ²⁰Northwestern University, Chicago, IL 60611, USA. ²¹National Institute of Mental Health, Bethesda, MD 20892, USA. ²²Brown University, Providence, RI 02912, USA. ²³University of Washington, Seattle, WA 98195, USA. ²⁴University of London, London, UK. ²⁵UCLA, Torrance, CA 90509, USA. ²⁶University of Michigan, Ann Arbor, MI 48109-2800, USA. ²⁷University of Utah, Salt Lake City, UT 84112, USA. ²⁸Banner Alzheimer's Institute, Phoenix, AZ 85006, USA. ²⁹UUC Irvine, Orange, CA 92868, USA. ³⁰Johns Hopkins University, Baltimore, MD 21205, USA. ³¹Richard Frank Consulting, New York, USA. ³²National Institute on Aging, Baltimore, MD, USA. ³³Oregon Health and Science University, Portland, OR 97239, USA. ³⁴University of Alabama, Birmingham, AL, USA. ³⁵Mount Sinai School of Medicine, New York, NY, USA. ³⁶Rush University Medical Center, Chicago, IL 60612, USA. ³⁷Baylor College of Medicine, Houston, TX, USA. ³⁸Wien Center, Miami Beach, FL 33140, USA. ³⁹Columbia University Medical Center, New York, NY, USA. ⁴⁰New York University, New York, NY, USA. ⁴¹University of Texas Southwestern Medical School, Galveston, TX 77555, USA. ⁴²Duke University Medical Center, Durham, NC, USA. ⁴³Emory University, Atlanta, GA 30307, USA. ⁴⁴Medical Center, University of Kansas, Kansas City, KS, USA. ⁴⁵University of Kentucky, Lexington, KY, USA. ⁴⁶Mayo Clinic, Jacksonville, FL, USA. ⁴⁷University of Rochester Medical Center, Rochester, NY 14642, USA. ⁴⁸Yale University School of Medicine, New Haven, CT, USA. ⁴⁹McGill Univ. Montreal-Jewish General Hospital, Montreal, PQ H3A 2A7, Canada. ⁵⁰Sunnybrook Health Sciences, Toronto, ON, Canada. ⁵¹U.B.C. Clinic for AD & Related Disorders, Vancouver, BC, Canada. ⁵²Cognitive Neurology - St. Joseph's, London, ON, Canada. ⁵³Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV 89106, USA. ⁵⁴Premiere Research Inst (Palm Beach Neurology), W Palm Beach, FL, USA. ⁵⁵Georgetown University Medical Center, Washington, DC 20007, USA. ⁵⁶Stanford University, Stanford, CA 94305, USA. ⁵⁷Boston University, Boston, MA, USA. ⁵⁸Howard University, Washington, DC 20059, USA. ⁵⁹Case Western Reserve University, Cleveland, OH 44106, USA. ⁶⁰Neurological Care of CNY, Liverpool, NY 13088, USA. ⁶¹St. Joseph's Health Care, London, ON N6A 4H1, Canada. ⁶²Dent Neurologic Institute, Amherst, NY 14226, USA. ⁶³Ohio State University, Columbus, OH 43210, USA. ⁶⁴Albany Medical College, Albany, NY 12208, USA. ⁶⁵Hartford Hospital Olin Neuropsychiatry Research Center, Hartford, CT 06114, USA. ⁶⁶Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA. ⁶⁷Wake Forest University Health Sciences, Winston-Salem, NC, USA. ⁶⁸Medical University South Carolina, Charleston, SC 29425, USA. ⁶⁹Nathan Kline Institute, Orangeburg, NY, USA. ⁷⁰University of Iowa College of Medicine, Iowa City, IA 52242, USA. ⁷¹USF Health Byrd Alzheimer's Institute, University of South Florida, Tampa, FL 33613, USA.