Original Research

The Journal of Prevention of Alzheimer's Disease - JPAD
Volume 8, Number 2, 2021

Molecular Subtyping of Mild Cognitive Impairment Based on Genetic Polymorphism and Gene Expression

H.-T. Li[1], S.-X. Yuan[1], J.-S. Wu[2], X.-Z. Zhang[3], Y. Liu[4], X. Sun[1] and For the Alzheimer's Disease Neuroimaging Initiative†

1. State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, P. R. China; 2. School of Geography and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing, P.R. China; 3. Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, USA; 4. The First Affiliated Hospital of Nanjing Medical University, Jiangsu Province Hospital, Nanjing, P.R. China

Corresponding Author: Xiao Sun, State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, P. R. China, xsun@seu.edu.cn

J Prev Alz Dis 2021;2(8):224-233
Published online November 23, 2020, http://dx.doi.org/10.14283/jpad.2020.65

Abstract

BACKGROUND: Alzheimer's Disease (AD) is a neurodegenerative brain disease in the elderly. Recent studies have revealed the heterogeneous nature of AD. Mild Cognitive Impairment (MCI) is the prodromal stage of AD.
OBJECTIVES: In this study, we identified subtypes of MCI based on genetic polymorphism and gene expression.
METHODS: We utilized the two types of omics data, namely genetic polymorphism and gene expression profiling, derived from 125 MCI patients' peripheral blood samples from the ADNI-1 dataset. Similarity network fusion (SNF) algorithm was implemented to cluster MCI patient subtypes. And 185 MCI patients in ADNI-2 were utilized to evaluate the effectiveness of this method. Two MCI subtypes were identified by implementing the SNF algorithm.
RESULTS: We used Kaplan-Meier analysis and log-rank testing for the conversion from MCI to AD between two subtypes, and p-value is 4.58×10^{-3}. In addition, we compared patients among two MCI subtypes by the following factors: the changes in Alzheimer's Disease cognitive scales and MRI image; significantly enriched pathways based on differentially expressed genes. This study proved that MCI is a heterogeneous disease by concluding that AD development in two MCI subtypes is significantly different.
CONCLUSIONS: MCI patients with different molecular characteristics have different risks converting to AD. In addition to evaluating statistics, genetic polymorphism and gene expression profiling from MCI patients' peripheral blood are non-invasiveness and cost-effectiveness markers to identify MCI subtypes for clinical application.

Key words: Alzheimer's disease, mild cognitive impairment, molecular subtyping, similarity network fusion.

Introduction

Alzheimer's disease (AD) is a chronic degenerative brain disease and the most common cause of dementia in the elderly. According to statistics, about 10% of people older than 65 suffer from AD (1). Due to the lack of understanding of its causes, effective drugs or treatments of AD is yet not invented.

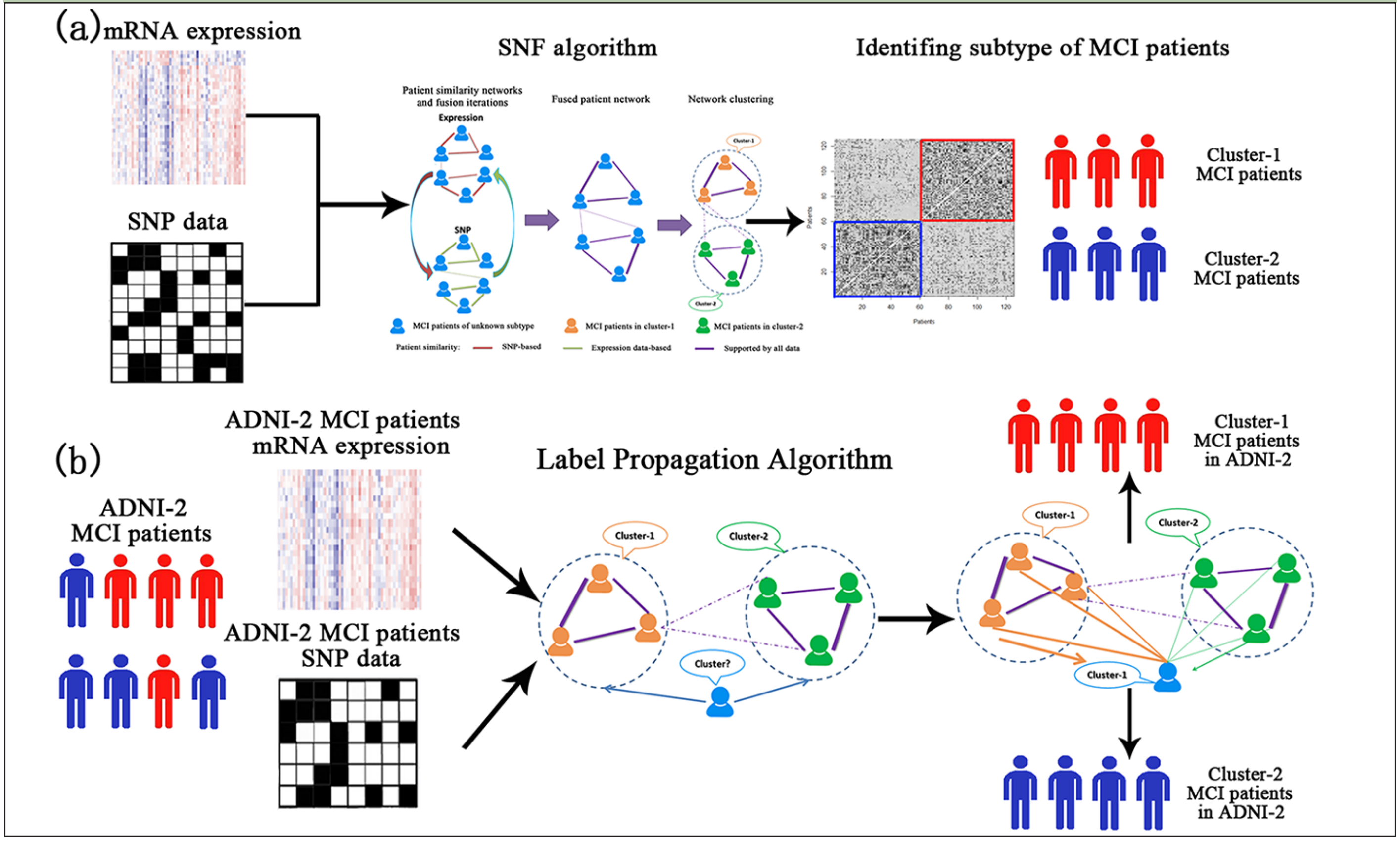
AD is a complex and heterogeneous disease caused by multiple different genetic factors (2). Recently, more and more studies, such as clinicopathologic (3), atrophy patterns on magnetic resonance imaging (MRI) (4) and amyloid-β fibril polymorphism on solid-state nuclear magnetic resonance (ssNMR) (5), have supported the hypothesis on the existence of distinctive AD molecular subtypes. For example, the rapidly progressive form in which neurodegeneration occurs within months and a typical prolonged-duration form are two AD clinical subtypes that been well recognized. Recently, some researchers have found that different AD clinical subtypes were correlated with fibril formations subtypes by researching on 37 brain samples from 18 deceased Alzheimer's patients obtained by using ssNMR (5). Lately, another research assigned 4,050 people with late-onset AD into six subgroups according to their cognitive functioning at the time of diagnosis and then utilized genetic data to find the biological differences across these subgroups (6). This study supported the biological coherence of cognitively defined subgroups. With more in-depth studies of Alzheimer's subtypes, new diagnostic criteria, and treatment of AD that target specific kinds of AD subtypes can be expected.

Mild Cognitive Impairment (MCI) is known as the prodromal stage of AD. MCI is a neurological disorder in which an elderly has mild but measurable changes in cognition. It is worth mentioning that not all people with MCI will develop AD. Studies suggest that MCI patients progress to AD at a rate of approximately 10% every year (7). Early identification of high-risk subtypes MCI patients appears to be significant and may enable a more effective, preventive treatment, thereby increasing the possibility of delaying even avoiding conversion from MCI to AD.

For the above reasons, we believe that MCI is a heterogeneous disease. Identifying the subtypes of MCI is critical for implementing precision medicine approaches

Received July 21, 2020
Accepted for publication October 20, 2020

Figure 1. Flow chart of our research. (a) The Similarity Network Fusion (SNF) algorithm is used to integrate SNP and gene expression data for subtype identification of MCI patients; (b) The label propagation algorithm is applied to predict the subtype of any new patient from ADNI-GO/2 for testing the effectiveness and reliability of the SNF algorithm



and for ultimately developing successful subtype-specific drugs for AD. And classifying MCI patients into meaningful subtypes may provide better targeted treatment to delaying or preventing the conversion from MCI to AD. Genetic factors play an important role in MCI and AD (2). However, to our knowledge, the molecular subtyping of MCI based on integrative multi-omic data was not taken into consideration among current studies. Therefore, in this study, we took advantage of the two types of omics data, including genetic polymorphism and gene expression, derived from 125 MCI patients' peripheral blood samples from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to identify the MCI patient subtypes (8). We used the Similarity Network Fusion (SNF) algorithm to cluster the two types of omics data to determine the subtypes of 125 MCI patients (9). For testing the effectiveness and reliability of the SNF algorithm, 185 MCI patients from ADNI-2 were identified the subtype by the label propagation algorithm (9, 10). The flow chart of our research is illustrated in Figure 1. To prove the biological and clinical significance of subtyping patients based on our method, these different subtypes were compared by the following factors: the time difference of the conversion from MCI to AD; cognitive scales and MRI image; significantly enriched pathways based on differentially expressed genes separately.

Methods

Genomic data and imaging data

Data used in this study were downloaded from ADNI. ADNI was a multi-site study proposed by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the Food and Drug Administration (FDA) in 2003. This organization is holding an ongoing, longitudinal, multicenter study. Its primary goal is to test whether clinical, imaging, genetic, and biochemical biomarkers are effective in clinical trials of MCI and AD. The first stage of ADNI, as known as ADNI-1, was completed in 2010 (8). More up-to-date and detail information is available at http://adni.loni.usc.edu/.

In this article, we used combinations of multi-omics data (genetic polymorphism and gene expression) from the ADNI-1 and ADNI-GO/2 study to identify the MCI molecular subtypes and to predict the conversion from MCI to AD. 125 MCI patients' SNP and gene expression data were downloaded from ADNI-1 for identification MCI subtypes. Meanwhile, 185 MCI patients were downloaded from ADNI-GO/2 as an independent verification dataset for predicting the subtype of any new patients. The information on MCI patients is listed in Supplementary Excel file 1. Both profiling were collected from peripheral blood samples. ADNI-1 and ADNI-GO/2 subjects were genotyped using the Human 610-Quad BeadChip and Illumina Human Omni Express BeadChip, respectively. Only SNP markers were analyzed for subsequent analysis. Quality control steps were performed on genetic polymorphism using the software package named PLINK (11), release v1.90b.5. SNPs with missing rate >0.05, minor allele frequency < 0.05, and Hardy–Weinberg equilibrium $P < 10^{-3}$ were excluded from the genetic polymorphism set. Then the SNP data was applied by using the IMPUTE2 program for imputing the missing data with NCBI 1000 Genomes build 37 (UCSC hg19) as the reference panel (12). The Affymetrix Human Genome U219 Array was carried out for expression profiling, which contains 530,467 probes. Thenceforth, we used an R package named RMA for the normalization of gene expression microarray data (13). Finally, 49,293 transcripts were kept in this study.

There are various clinical/cognitive assessment scores from ADNI that are useful to compare clinical information between two subtypes of patients, including Mini Mental State Examination (MMSE), Clinical Dementia Rating Sum of Boxes (CDR-SB) and Activities of Daily Living Score (from the Functional Activities Questionnaire, FAQ). In addition, we downloaded T1 weighted MRI images in NIFTI format from 125 MCI patients' baseline, 24-month follow-up data set in ADNI, and structural MRI scan applied inversion recovery-fast spoiled gradient recalled (IR-SPGR) for researching two clusters of MCI patients' differences in areas of brain atrophy. VBM analyses were performed using the SPM12 toolkit (Statistical Parametric Mapping software, http://www.fil.ion.ucl.ac.uk/spm/sofware/ spm12) running under MATLAB 2013a (14).

MCI subtype identification based on similarity network fusion

We applied the similarity network fusion (SNF) algorithm to cluster the MCI patient subtypes (9). SNF is an integrated characterization of genomic profiling at multiple levels for subtype identification. The advantage of using SNF is that it is based on complementarity in multiple genomic data types. First, the SNF algorism uses a similarity measure to constructs a patient-by-patient similarity network for each genomic data type. The nodes of the network for each data type represent patients and the weighted edges are equivalent to pairwise sample similarities. Next, the network fusion step updates every network using a nonlinear method named message-passing theory. Each iteration makes these networks more similar to each other. After many iterations, multiple networks converge to a fusion network. Finally, the fusion network is clustered into several subtypes based on spectral clustering methods. The illustrative example of SNF steps are shown in Figure S1. Some patients (002_S_0729, 010_S_0161 and 011_S_1282 from cluster-1; while 005_S_0546, 027_S_1045 and 037_S_0150 from cluster-2) were used as examples to explain the clustering process of the SNF method (Figure S1 (d)).

More formally speaking, given n MCI patients and M omics (SNP and expression data in this study), the sample×sample similarity graph G=(N, W) is constructed, where node set N represents the samples $x_1, x_2, \ldots, x_n$ and the edges weight W(i, j) represents the weight between x_i and x_j. W is defined by:

$$W = exp\,(-\frac{d^{(m)}(x_i, x_j)}{\alpha \varepsilon_{i,j}})$$

where $d^{(m)}$ (x_i, x_j) is the Euclidean distance between sample x_i, x_j for the m-th omic. α is a hyperparameter and α=0.8 in this study. ε is expressed as below:

$$\varepsilon_{i,j} = \frac{mean(\varepsilon(x_i, K_i)) + mean\left(\varepsilon(x_j, K_j)\right) + \varepsilon(x_i, x_j)}{3}$$

where K_i is the number of neighbours of x_i and K_i=30 in this study, mean (ε(x_i, K_i)) is the average distance between x_i and each of its neighbors. ε is introduced to eliminate the scaling problem.

A transition probability matrix is constructed between all MCI patients initially by:

$$P_1^{(m)}(i,j) = \begin{cases} \frac{W^{(m)}(i,j)}{2\sum_{k\neq i} W^{(m)}(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases}$$

Meanwhile, a transition probability matrix between nearest neighbors is defined by:

$$S^{(m)}(i,j) = \begin{cases} \frac{W^{(m)}(i,j)}{\sum_{k\neq i} W^{(m)}(i,k)}, & j \in N_i \\ 0, & otherwise \end{cases}$$

where N_i represent a set of i's k nearest neighbors in matrices with measurements from the m-th omic.

Then, the matrix P is updated based on message-passing theory iteratively between the k nearest neighbors by formula:

$$P_{q+1}^{(m)} = S^{(m)} \frac{\sum_{k\neq m} P_q^{(k)}}{M-1} S^{(m)^q},$$

where $P_q^{(m)}$ is the matrix for omic m at iteration q. The iterative process means that the connection information of different networks is exchanged to achieve the final uniform network.

After completing the network fusion, low-weight edges in each network disappear, and high-weight edges are retained. SNF reduces the noise among these steps, which makes this method robust to noise and the data heterogeneity. Finally, based on spectral clustering methods, namely minimize RatioCut, the fusion network is clustered into several subgroups. Such subgroups are considered as our resulting subtypes. The details of SNF reference (9).

Any new MCI patient' subtype prediction based on label propagation

We adopted label propagation algorithm which is a simple iterative semi-supervised learning algorithm based on network structure to identify the subtype of the new MCI patient (9, 10). Assume n patients have been determined into y subtypes by the SNF method with a fused network F. To predict the subtype of a new patient, a similarity matrix F=[F s;s' 1] is constructed, where s is the similarities vector calculated by SNF. Define a (n+1)×(n+1) probabilistic transition matrix T:

$$T_{ij} = \frac{\hat{F}_{ij}}{\sum_{k=1}^{n+1} \hat{F}_{kj}}$$

where T_{ij} is the probability of jumping from node j to i. Also we define a (n+1)×y label matrix Y, whose i-th row representing the label probabilities of node y_i. We iterate the propagation process as follows:

Repeat the following steps:

$$Y_{t+1} = \hat{F} * Y_t$$

$$Y_{t+1}(1{:}n) = y$$

This process will converge usually in 1000 iterations. And we can predict the subtype of the new patient given by converged Y.

Results

Clustering of MCI patients

We downloaded 138 MCI patients' gene expression profiling and 361 MCI patients' genetic polymorphism data from the ADNI-1 dataset. The number of MCI patients with both genetic polymorphism and gene expression was 125. Hence, we used these MCI patients in this article for integrating the two types of omics data to identify MCI patient subtypes. Moreover, 276 MCI patients' SNP data and 302 gene expression profiling were downloaded from the ADNI-GO/2 dataset. 185 MCI patients who have SNP data, gene expression data, and clinical follow-up data for greater than 36 months were selected as an independent verification set to evaluate the effectiveness of this method. Table S1 shows the characteristics of the MCI patients included in this study.

The subtypes of MCI patients in the ADNI-1 dataset were identification based on SNF method (9). In the beginning, quality control steps were performed on genetic polymorphism using the software package named PLINK (11) and gene expression profiling using an R package named RMA (13) as described in method. Then, we utilized SNF to cluster MCI patients using both SNP and gene expression profiling after quality control. SNFtool R package (v2.3.0) was applied with the parameters K = 30, alpha = 0.8, T = 20 (9). Spectral clustering implemented in the SNFtool package was run on the SNF fused similarity matrix to obtain the groups that each corresponding to k=2 to 5.

After executing the SNF algorithm, we chose the best number of clusters according to two main approaches of the spectral clustering method. One is the connectivity of the network, and the other is to make use of the structure of eigenvectors of the Laplacian L (9). However, the optimal number of clusters based on the connectivity of the network is 2, the best number decided by the other approaches is 3. Therefore, we used the highest average silhouette score as an assistance approach to decide the optimal number of clusters. The silhouette score represents the coherence of clusters to evaluate whether patients are more similar within subtypes. In other words, the silhouette score condenses the cluster quality for each patient's omics data into a single score that ranges from 1.0 to -1.0. Hence, we had identified two subtypes. The number of patients in cluster-1 is 61, and cluster-2 has 64 patients.

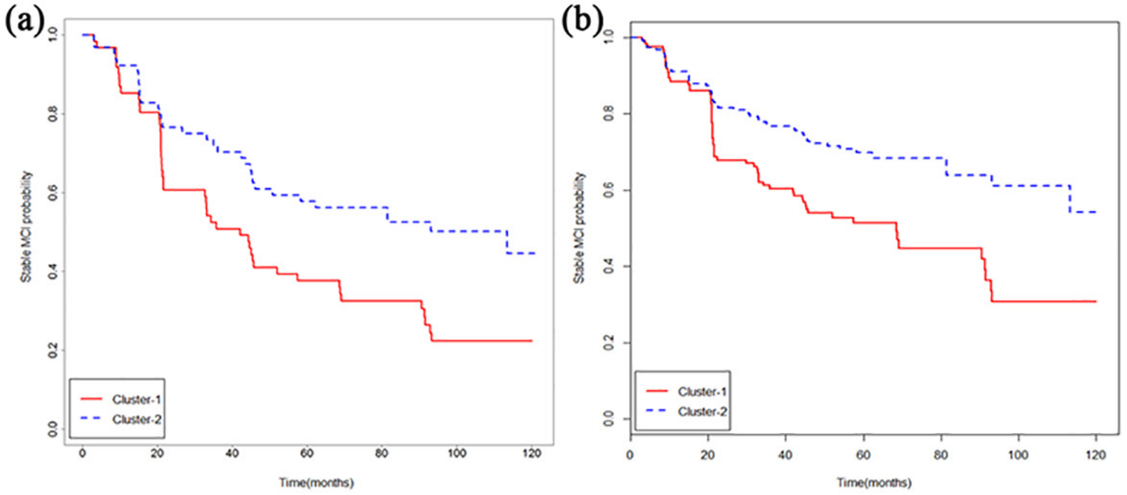
To prove the biological and clinical significance of subtyping patients based on the SNF method, we applied the label propagation algorithm to assign new patients to subtypes in the ADNI-2 datasets (9, 10). Genotype data of MCI patients from the ADNI-GO/2 dataset were downloaded, quality controlled, imputed to the Illumina 610Quad platform and combined. Genotype imputation was conducted to estimate unobserved genotypes. Impute2 software was used with NCBI 1000 Genomes build 37 (UCSC hg19) as the reference panel (12). After executing the label propagation algorithm to 185 patients in ADNI-GO/2, 60 MCI patients were identified in cluster-1, while 125 patients were identified in cluster-2. The detail information on the subtypes of MCI patients in the ADNI-1 and ADNI-GO/2 dataset is listed in Supplementary Excel file 1.

Two MCI subtypes supported by clinical manifestations

We first examined the time difference of the conversion from MCI to AD between two subtypes of patients. Because the exact date of conversion to AD was not known, we used the midpoint between the last follow-up without an AD diagnosis and the first follow-up with an AD diagnosis for analyses. Subjects who did not convert were censored at the time of their last interview. We performed a Kaplan-Meier analysis on MCI of these two clusters. As is shown in Figure 2(a), P-value is 4.58×10^{-3}, demonstrating a significantly different amount of time is consumed for MCI-to-AD conversion between two clusters. Patients that develop the disease more rapidly (red solid line) were cluster-1 MCI patients, and the others (blue dashed line) were cluster-2 MCI patients.

We also considered the changes in Alzheimer's Disease cognitive scales. Cognitive function status was measured

Figure 2. The Kaplan-Meier plot analysis on MCI of the two clusters of clinical data. X axis represents time past after MCI patients participating the study, while Y axis represents estimated percentages of stable MCI patients. The red solid line represents cluster-1 MCI patients in ADNI-1 (a) and ADNI-GO/2 (b), while the blue dashed line represents cluster-2 MCI patients in ADNI-1 (a) and ADNI-GO/2 (b)



by the Mini-Mental State Examination (MMSE) (rating 0–30, higher scores indicate good cognitive function), the Clinical Dementia Rating Sum of Boxes (CDR-SB) (rating 0–25, with higher scores representing greater impairmen) and the Functional Assessment Questionnaire (FAQ) (range 0–30, with higher scores representing greater impairment) in two years for two MCI subtypes of patients (8). As is shown in Figure 3(a), cognitive decline in cluster-1 MCI patients tends to be more remarkable than that of cluster-2 over 24 months.

To test the effectiveness and reliability of the SNF algorithm through its application on ADNI-GO/2 patients, we examined the time difference of the conversion from MCI to AD between two subtypes of all MCI patients. As is shown in Figure 2(b), this gives a log-rank P-value of 2.26×10^{-4}. And three AD cognitive scales were also displayed in two years for two MCI subtypes of patients in the ADNI-2 dataset, which is shown in Figure 3(b). The scores change trends of all three cognitive scales in ADNI-GO/2 are similar to the ADNI-1 dataset. Thus, it proved the validity of the SNF method for subtyping MCI patients based on integrative genetic polymorphism and gene expression. Meanwhile, the cluster-1 subtypes having the worse prognosis than the cluster-2 subtypes.

Two MCI subtypes supported by MRI image

We further analyzed the MRI images to illustrate the difference between two clusters of MCI patients' ADNI baseline and 24-month follow-up MRI dataset using voxel-based morphometry (VBM) analyses in atrophy areas (15). VBM analysis has been developed for characterizing differences in the local composition of brain tissue using MRI and is not restricted to previously called region-of-interest measurements.

Firstly, we normalized images with the voxel sizes of 1.5×1.5×1.5mm^3 because it could preserve the total amount of signal in the images. After normalization, T1-weighted images were segmented into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) using default option parameters on SPM12's unified segmentation procedure. After that, we transformed patients' images to the Montreal Neurological Institute (MNI) co-ordinate space using a template. Cognitive impairment is related to the MRI of GM decline on longitudinal analysis. Hence, on GM images, the spatial normalization approach was performed with the diffeomorphic anatomical registration using the exponentiated Lie algebra (DARTEL) algorithm (16). Subsequently, the images were smoothed with a 10-mm full-width at half-maximum isotropic Gaussian smoothing kernel. The results of GM images were analyzed with the two-sample t-test. For voxels in GM probability maps between baseline and 24 months, we selected those voxels with P<0.05 corrected by False Discovery Rate (FDR), and only regions of more than 100 contiguous selected voxels were considered in the analysis. To analysis the result of GM atrophy origins, we utilized the predefined anatomical masks obtained from an extension to the SPM package – XjView toolbox (http://www.alivelearn.net/xjview/) and the automated anatomical labeling (AAL, http://www.gin.cnrs.fr/en/tools/aal-aal2/) (17).

Based on the current official anatomical nomenclature proposed by Guilherme et al., the brain structure was divided into six lobes: frontal, parietal, occipital, temporal, insular, and limbic (18). The atrophic number of significantly different voxel regions is shown in Table 1. The result of the above steps was characterized by XjView.The comparison of cluster-1 (a) and cluster-2 (b) MCI patients' regions of gray matter atrophy between baseline and 24-month follow-up MRI images are shown in Figure 4(a,b).

Figure 4(c) reveals that the atrophic size of significantly

Figure 3. Changes in AD cognitive scales (MMSE, CDR, FAQ) in two years for two MCI subtypes in ADNI-1 (a) and ADNI-GO/2 (b). X axis represents time past after MCI patients participating the study, while Y axis represents Alzheimer's Disease cognitive scales score. Cognitive decline in cluster-1 MCI patients (red) is tend to be more remarkable than that of cluster-2 (blue) over 24 months

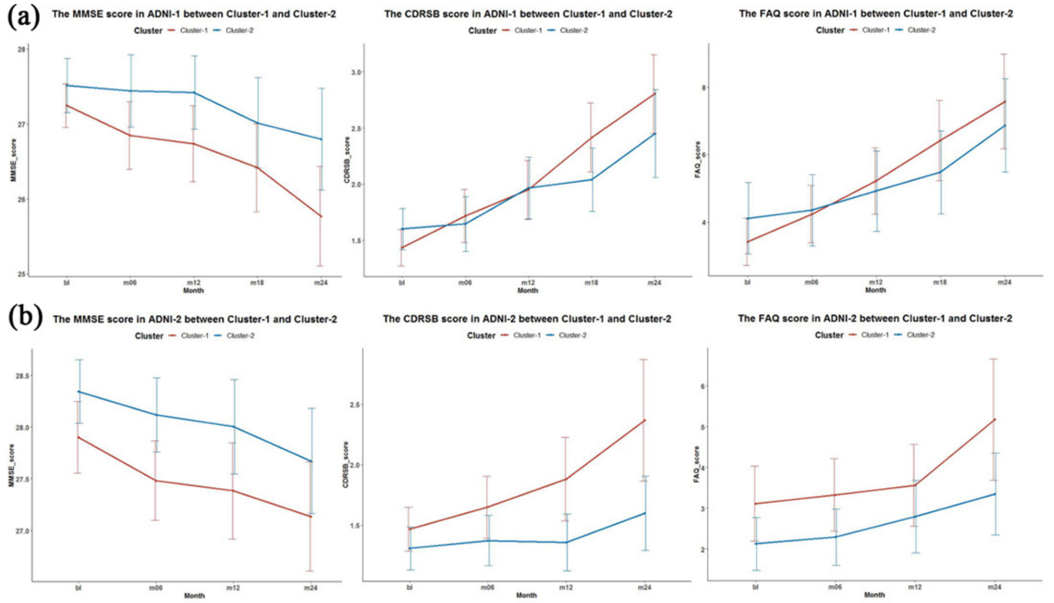
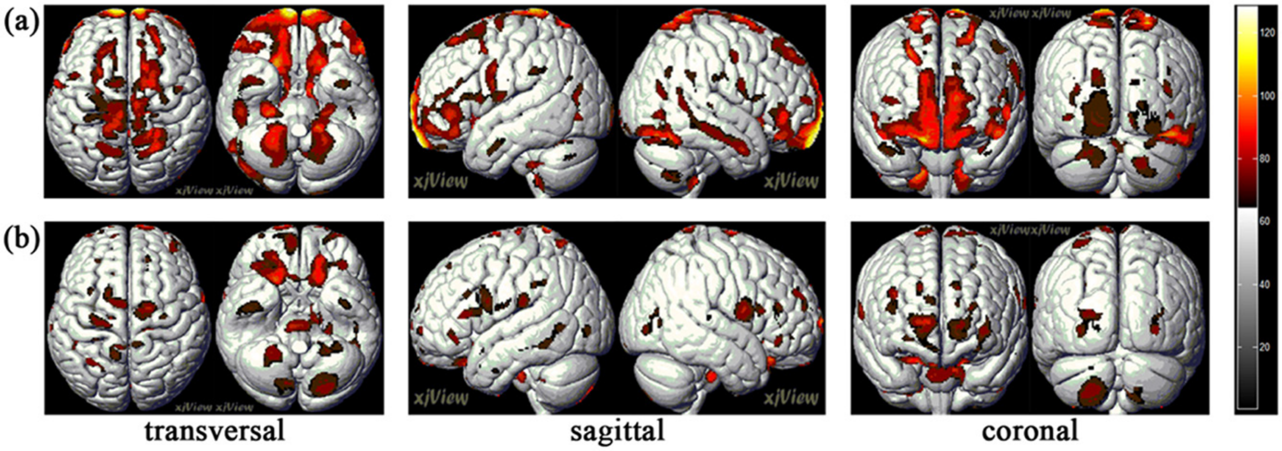


Figure 4. . Display of voxels with significantly brain areas of decreased gray matter intensity in each cluster. Images are 3D render view of (a) cluster-1 and (b) cluster-2 in sagittal, coronal and transversal. And paired images are MCI patients' baseline MRI images compared to those of 24-month follow-up using VBM analyses. Colored voxels show regions that were significant in the analyses with p<0.05 corrected by FDR, and regions threshold of 100 contiguous voxels. The color brighter (yellow) indicates the more significant area of brain atrophic voxels in 24 month. (c) The atrophic size of significantly different voxel bunches within six lobes in 24 months

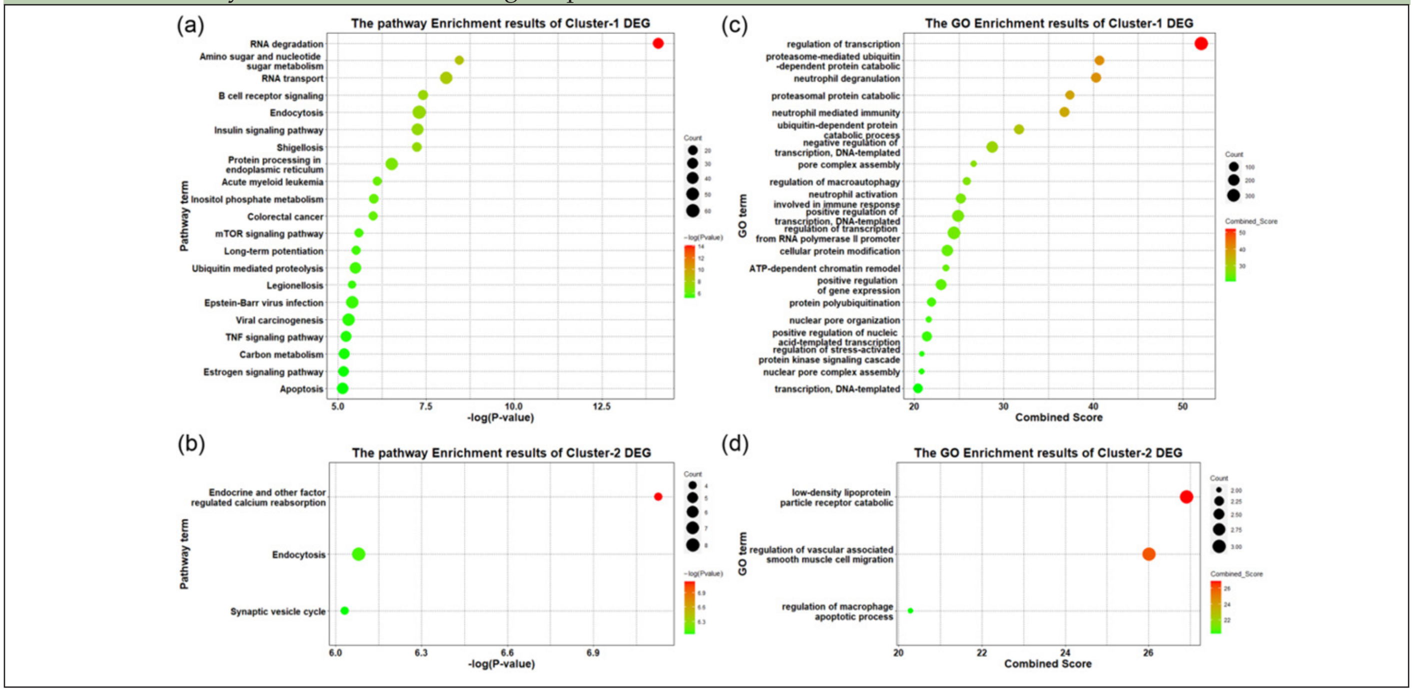


(c)

	Total Number of atrophic voxels	Number of atrophic voxels in six lobes	Frontal Lobe	Parietal Lobe	Occipital Lobe	Temporal Lobe	Limbic Lobe	insular
Cluster 1	83795	38761	21508	1804	5029	5110	4496	814
Cluster 2	26279	6571	4480	258	779	501	386	167

different voxel bunches of cluster-1 MCI patients in 24 months are apparently larger than that of cluster-2 MCI patients. In addition, the proportion of the atrophic voxels in six lobes accounted for 46.26% of total number of brain atrophic voxels in cluster-1, while in cluster-2 this ratio is 25.00%. This result indicates that not only was the atrophy of voxels in cluster-1 patients significantly more than that of cluster-2 patients, but also the location

Figure 5. The enriched significant KEGG pathways and GO biological processes bubble plot of differentially expressed genes (DEG) with FDR<0.05. (a) cluster-1 DEG KEGG enrichment, (b) cluster-2 DEG KEGG enrichment, (c) cluster-1 DEG GO enrichment, (d) cluster-2 DEG GO enrichment. The size of the dots represents the count of DEG in the corresponding pathways or GO terms. Y axis represents the enrichment pathways and biological processes. (a, b) X axis represents the opposite of the logarithm of p-value for each pathway, and (c, d) X axis represents the combine score which is defined by Enrichr for each biological process



of atrophy was also concentrated in the functional areas of the brain. Therefore, by comparing the MRI images of cluster-1 and cluster-2 MCI patients collected from two-year data, one can see that AD development of cluster-1 patient is faster than that of cluster-2. Hence, this proves the usefulness of the subtype classification in clinical.

Two MCI subtypes supported by gene annotation

Subsequently, differential expressions of mRNA of MCI cluster-1, cluster-2 compared with the cognitively normal samples were each computed using R package named limma (19). Adjust-P value< 0.05 served as the screening conditions for the significant differences. The significantly different expression gene-set of cluster-1 had 3156 genes, while that of cluster-2 had 178 genes. We applied the functional annotation tool of "Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment" and "Gene ontology (GO)" in Enrichr, which was an integrative web-based software application that included many new gene-set libraries for gene and sequence annotations (20). Enrichr provided an adjustment P-value and combined score to annotate the biological significance of differentially expressed genes. An adjust p-value, 0.1, was chosen as significant thresholds upon filtering the pathway data. Because of too many biological processes in GO, a threshold of the combined score was considered. Common GO analyses were performed with a cut-off of combined score 20 and adjust p-value 0.1. The definition of the combined score in EnrichR is to integrate both p-value and z-score with the formula c = z-score•log(p-value), where c is the combined score, represented by p-value computed using the Fisher exact test, and z-score computed by assessing the deviation from the expected rank. The significant pathways and biological processes of differentially expressed genes between cluster-1, cluster-2, and control are shown in Figure 5. The full list of KEGG pathways and GO enrichment analysis information is in Supplementary Excel file 2.

The most remarkable pathways of cluster-1 are the following: RNA degradation, Amino sugar and nucleotide sugar metabolism, and RNA transport. These pathways are related to a wild range of biological processes. Meanwhile, the significant pathways of cluster-1 were predominated by immune system-related biological fields, such as B cell receptor signaling, TNF signaling and some microbial infection pathway (Epstein-Barr virus infection, Shigellosis and Legionellosis). More research results showed that inflammation is closely related to AD. In the brain, immune system cells called microglia is activated by the presence of toxic amyloid-β and tau proteins (21). Microglia tries to get rid of the remnants of inflammasomes in tiny clumps. However, these remnants continued to spread new amyloid-β clusters as well as aggravating the state of AD. Notably, Epstein-Barr virus infection is also one of the significant pathways. Epstein-Barr virus is known to be the one of herpes viruses. Recent research indicated that herpes viruses abundance was significantly associated with modulators of APP metabolism which revealed viral regulation of AD risk by multiscale networks (22). And insulin signaling pathways have a close relationship with AD. AD has been considered as a metabolic dysfunction disease associated with impaired insulin signaling (23). Proteolytic processes contribute to the amyloid cascade, and proteolysis of tau may be critical to neurofibrillary degeneration, which correlates with AD (24).

The significant biological processes in GO enrichment of cluster-1 are the regulation of transcription and protein catabolic process. These biological processes are closely correlated. Regulation of the transcription, DNA-templated is any process that modulates the frequency, rate or extent of cellular DNA-templated transcription. Regulation of transcription of AD genes might be an important player in the neurodegenerative process. For example, the APP gene is ubiquitously expressed in a variety of tissues, with the highest expression shown in neuronal cells. The abnormally expressed APP will lead to an increased amount and deposition of the amyloid β peptide (Aβ) in the brain triggering AD-related neuronal degeneration (25). Mutant forms of ubiquitin may inhibit proteolysis within neurons, making these cells susceptible to inclusion formation. Therefore, some researchers hold the contention that neurodegenerative diseases collectively referred to as "ubiquitin protein catabolic disorders". Especially, similar to the KEGG analysis of cluster-1 MCI patients, the significant biological processes are also associated with the immune system. For instance, some biological processes are related to neutrophil and macroautophagy (26). Neutrophils are key components for early innate immunity. Blood samples from AD patients with dementia revealed that the neutrophil hyperactivation was associated with increased reactive oxygen species production as well as the levels of intravascular neutrophil extravascular traps. Moreover, neutrophil phenotype may have a close relationship with the rate of cognitive decline (26).

The cluster-2 significantly enriched pathways mainly consisted of neuronal signaling-related pathways, such as endocytosis and synaptic vesicle cycle. For instance, the synaptic vesicle cycle plays an important role in the biological process of exocytosis and endocytosis. It facilitates a series of events achieving chemical neurotransmission between functionally related neurons. Some study results demonstrated that considerable changes in the expression and functions of presynaptic proteins attributed in parts to direct effects of amyloid-β production and toxicity on the synaptic vesicle cycle (27). In addition, endocytosis is critical for the normal processing of APP, which is central to AD pathogenesis (28).

The most remarkable GO biological process of cluster-2 is the regulation of vascular associated smooth muscle cell migration. The degenerated smooth muscle cells express increased amounts of amyloid β-precursor protein deposition in the medial layer of the cerebral vessel wall and produce Aβ peptide (29). And the low-density lipoprotein particle receptor catabolic process is another important biological process in cluster-2. This biological process results in the breakdown of a low-density lipoprotein particle receptor molecule, a macromolecule that undergoes combination with a neurotransmitter to initiate a change in cell function. The disorder in this biological process could impair the neurotransmitter-triggered signal transduction appearing in AD.

Discussion

AD is a neurodegenerative brain disease that yet has no available effective medications or supplemental treatment. Studies have shown that AD is a heterogeneous disease. In this article, we integrated two types of omics data (genetic polymorphism and gene expression profiling) of MCI patients to identify subtypes with biological and clinical significance by the SNF method. We performed SNF, the integrative clustering of multiple genomic data algorithms, to cluster MCI patients. Experimental studies were conducted on subtypes of MCI patients, and we showed that multi-omics data define subtypes characterized by biological and clinical significance.

We utilized the SNF method to identify MCI patient subtypes based on multi-omics characteristics (9). SNF has been used to cluster subtype of specific cancer patients, and satisfactory results have been achieved. After executing the SNF algorithm, we identified two MCI subtypes. By comparing clinical information between two subtypes of patients, we considered the changes in two years on AD cognitive scales (MMSE, CDR, and FAQ) and MRI images in atrophy areas based on VBM. We found that the molecular subtypes of MCI are remarkably different in clinical information. It is necessary to lay the foundation for the precision treatment of MCI patients.

To study the difference in the disease mechanism of cluster-1 and cluster-2, differential expressions of MCI cluster-1, cluster-2 mRNA compared with the cognitively normal samples were computed correspondingly. And the differential expression genes in cluster-1 are significantly more than that of cluster-2. We conjecture that the risk factors of AD in cluster-1 are more complicated. Subsequently, we applied the functional annotation tool of KEGG and GO in Enrichr for enrichment analysis based on these genes. In cluster-1 MCI patients, there are some microorganisms (such as gram-negative bacterium and herpes viruses (22)) that can escape immune responses. These microorganisms activated immune responses, such as microglia, to clear the toxic proteins and widespread remnants from dying cells. Furthermore, these remnants continue to spread new amyloid-β clusters causing inflammatory storms (21). Above is the reason that MCI in cluster-2 patients may have synaptic failure and degeneration conditions. For example, the reduction in synaptic vesicle proteins has been shown to have a strong association with the clinical symptoms of dementia (27). We speculated that it is the storm caused by inflammasomes in the brain that result in cluster-1 MCI patients to develop the disease more rapidly than cluster-2 patients. Also, the perturbations of many other pathways have associated with the cause of AD. For example, Moriguchi et al. proposed that AD may be brain diabetes, and insulin signaling pathway is an important pathway for causing AD (23). And perturbation of pathways such as protein processing in endoplasmic reticulum, inositol phosphate metabolism and fubiquitin mediated proteolysis pathways will contribute to the amyloid cascade, which closely related to senile plaques and thus causing AD (24). The cluster-2 significantly enriched pathways mainly consisted of neuronal signaling-related pathways, and some scholars considered AD as a synaptic dysfunction caused by diffusible oligomeric assemblies of the amyloid-β protein (27). Both cluster-1 and cluster-2 enriched KEGG pathways of significantly differentially expressed genes have the endocytosis pathway. Hence, we speculated that endocytosis is the basic molecular mechanism of AD.

SNP data and mRNA expression profiling collected from patients' peripheral blood have the characteristics of non-invasiveness and cost-effectiveness markers to identify MCI subtypes for clinical application. Clinical decisions will most likely be dictated by the genetic characteristics of AD patients in the coming years. We believe our method can effectively identify the subtypes of MCI patients, and can be applied in clinical in the future. Tailoring our method based on individual genetic characteristics will help doctors and researchers develop better therapeutic strategies and save many of MCI patients from receiving unnecessary toxic therapy. Further study should take into account the factors that can influence gene expression. For example, some other pathologies, influencing the expression of certain genes, may be present in elderly MCI patients. It may have an impact on the subtyping of MCI patients.

Two experiments can illustrate the clinical relevance of our method. For the first experiment, the expression data of 44 AD patients at baseline from the ADNI dataset were downloaded. We performed a hierarchical clustering analysis of patients with AD and patients of the two subtypes of MCI based on expression data using a similarity measure in SNF. The results are shown in the following Figure S2. This figure clearly shows that most AD patients are clustered with MCI cluster-1 patients. For the other experiment, 27 patients with AD at baseline in the ADNI dataset were downloaded. We applied the label propagation algorithm to assign new patients to subtypes. The subtype labels of these MCI patients were listed in Table S2. To test the effectiveness and reliability of our method, three AD cognitive scales were also displayed in 24-month for two subtypes of AD patients. As is shown in Figure S3, cognitive decline in cluster-1 MCI patients tends to be more remarkable than that of cluster-2 over 24 months, which is similar to the MCI patients in the ADNI dataset.

Hence, we believe our method can effectively identify the subtypes of MCI patients, and can be applied in clinical in the future. We look forward to potential collaborations with doctors and experimental biologists. We hope that the subtyping of MCI patients predicted with our model, will demonstrate its medical and therapeutic meaning. Besides, different types of data share complementary information, which is robust to noise and data heterogeneity (9). In the future, other types of biological data, such as DNA methylation and miRNA expression, can be integrated to explore biological patterns related to identify MCI subtypes. And classifying MCI patients into meaningful subtypes may improve the forecasting performance to proposing a method for predicting the conversion from MCI to AD (30).

Availability of data and material: Data used in this study are available through the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu).

Author Contributions: HTL, SXY were involved with conception, design, and interpretation of data. HTL and JSW were involved with data analysis. XS, YL and XZZ provided general overall supervision of the study. XS acquired funding. All authors contributed to drafting and critical revision of the manuscript and have given final approval of the version to be published.

Funding: This research was sponsored by the National Natural Science Foundation of China (61972084, 81830053) and the Key Research and Development Program of Jiangsu province of China (BE2016002-3).

Acknowledgements: Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). 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. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical Standards: Data involved in this study came from the Alzheimer's disease neuroimaging initiative (ADNI) database. All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the Helsinki Declaration of 1975.

References

1. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology 2013;80:1778–1783
2. Bagyinszky E, Youn YC, An SSA, Kim S. The genetics of Alzheimer's disease. Clin Interv Aging 2014;9:535–551
3. 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;10:785–796
4. Park J-Y, Na HK, Kim S, et al. Robust Identification of Alzheimer's Disease subtypes based on cortical atrophy patterns. Sci Rep 2017;7:43270
5. Qiang W, Yau W-M, Lu J-X, Collinge J, Tycko R. Structural variation in amyloid-β fibrils from Alzheimer's disease clinical subtypes. Nature

2017;541:217–221

6. Mukherjee S, Mez J, Trittschuh EH, et al. Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. Molecular psychiatry 2018
7. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand 2009;119:252–265
8. Weiner MW, Aisen PS, Jack CR, et al. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. Alzheimers Dement 2010;6:202–11.e7
9. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11:333–337
10. Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation. 2002
11. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 2007;81:559–575
12. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44:955–959
13. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15
14. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. Neuroimage 2000;11:805–821
15. Wright IC, McGuire PK, Poline JB, et al. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 1995;2:244–252
16. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage 2007;38:95–113
17. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 2002;15:273–289
18. Ribas GC. The cerebral sulci and gyri. Neurosurg Focus 2010;28:E2
19. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47
20. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44:W90-97
21. Abbott A. Is 'friendly fire' in the brain provoking Alzheimer's disease? Nature 2018;556:426
22. Readhead B, Haure-Mirande J-V, Funk CC, et al. Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus. Neuron 2018;99:64-82.e7
23. Moriguchi S, Ishizuka T, Yabuki Y, et al. Blockade of the KATP channel Kir6.2 by memantine represents a novel mechanism relevant to Alzheimer's disease therapy. Molecular Psychiatry 2018;23:211–221
24. Dickson DW. Apoptotic mechanisms in Alzheimer neurofibrillary degeneration: cause or effect? J Clin Invest 2004;114:23–27
25. Theuns J, Van Broeckhoven C. Transcriptional regulation of Alzheimer's disease genes: implications for susceptibility. Hum Mol Genet 2000;9:2383–2394
26. Dong Y, Lagarde J, Xicota L, et al. Neutrophil hyperactivation correlates with Alzheimer's disease progression. Ann Neurol 2018;83:387–405
27. Ovsepian SV, O'Leary VB, Zaborszky L, Ntziachristos V, Dolly JO. Synaptic vesicle cycle and amyloid β: Biting the hand that feeds. Alzheimer's & Dementia 2018;14:502–513
28. Karch CM, Goate AM. Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis. Biological Psychiatry 2015;77:43–51
29. Van Nostrand WE, Melchor J, Wagner M, Davis J. Cerebrovascular smooth muscle cell surface fibrillar A beta. Alteration of the proteolytic environment in the cerebral vessel wall. Ann N Y Acad Sci 2000;903:89–96
30. Vinutha N, Pattar S, Sharma S, Shenoy PD, Venugopal KR. A Machine Learning Framework for Assessment of Cognitive and Functional Impairments in Alzheimer's Disease: Data Preprocessing and Analysis. J Prev Alzheimers Dis 2020;7:87–94