Integration of Imaging Genomics Data for the Study of Alzheimer's Disease Using Joint-Connectivity-Based Sparse Nonnegative Matrix Factorization

Kai Wei (✉ 201930310010@stu.shmtu.edu.cn)

Shanghai Maritime University <https://orcid.org/0000-0003-4703-8041>

Wei Kong

Shanghai Maritime University

Shuaiqun Wang

Shanghai Maritime University

Research Article

Keywords: Index Terms-Alzheimer's Disease (AD), Imaging Genetics, Nonnegative Matrix Factorization (NMF)

Posted Date: July 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-586682/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Journal of Molecular Neuroscience on August 19th, 2021. See the published version at <https://doi.org/10.1007/s12031-021-01888-6>.

Integration of Imaging Genomics Data for the Study of Alzheimer's Disease Using Joint-Connectivity-Based Sparse Nonnegative Matrix Factorization

Kai Wei^{1†}, Wei Kong^{1†}, Shuaiqun Wang¹

Kai Wei and Wei Kong should be regarded as joint first authors.

¹College of Information Engineering, Shanghai Maritime University, 1550 Haigang Ave., Shanghai 201306, P. R. China

Corresponding author: Wei Kong (e-mail: weikong@shmtu.edu.cn).

Abstract

Image genetics reveal the connection between microscopic genetics and macroscopic imaging and then detect diseases' biomarkers. In this research, we make full use of the prior knowledge that has significant reference value for exploring the correlation of brain and genetics to explore more biologically substantial biomarkers. In this paper, we propose Joint-Connectivity-Based Sparse Nonnegative Matrix Factorization (JCB-SNMF). The algorithm simultaneously projects structural Magnetic Resonance Imaging (sMRI), Single Nucleotide Polymorphism sites (SNPs), and gene expression data onto a common feature space, where heterogeneous variables with large coefficients in the same projection direction form a common module. In addition, the connectivity information of each region of the brain and genetic data are added as prior knowledge to identify regions of interest (ROI), SNP, and gene-related risks related to Alzheimer's disease (AD) patients. GraphNet regularization increases the anti-noise performance of the algorithm and the biological interpretability of the results. The simulation results show that compared with other NMF-based algorithms (JNMF, JSNMF), JCB-SNMF has better anti-noise performance and can identify and predict biomarkers closely related to AD from significant modules. By constructing a protein-protein interaction network (PPI), we identified SF3B1, RPS20, and RBM14 as potential biomarkers of AD. We also find some significant SNP-ROI and Gene-ROI pairs. Among them, two SNPs of rs4472239 and rs11918049 and three genes of KLHL8, ZC3H11A, and OSGEPL1 may have effects on the gray matter volume of multiple brain regions. This model provides a new way to further integrate multimodal impact genetic data to identify complex disease association patterns.

Index Terms-Alzheimer's Disease (AD), Imaging Genetics, Nonnegative Matrix Factorization (NMF)

1. Introduction

Imaging genetics has been widely used in neurodegenerative diseases. It can explore the influence of genes on brain structure and function and use brain imaging to evaluate the impact of

genes on individuals. Recently, it has made significant progress in studying the pathogenesis of AD and mining AD-related biomarkers.

Canonical Correlation Analysis (CCA) [1] is an effective method that integrates two or more different modal data types. It can maximize the linear combination of the most remarkable correlation among different types of variables and then obtain the data's interrelated components. However, due to the high-dimensional characteristics of image genetics, the accurate association analysis of the different modal data is challenged in the case of limited samples. Many scholars added various sparse constraints based on CCA to prevent the model from overfitting [2][3][4].

Unfortunately, the CCA-based model can only find a pair of canonical correlation variables simultaneously [5]. It is possible to limit the number of features for the same type of data. Introducing the non-negative matrix factorization (NMF) method into the field of image genetics can better overcome these problems.

NMF is a robust dimensionality reduction computing framework that can integrate different omics data. Specifically, NMF can decompose a non-negative matrix into two matrices: the base matrix W and the coefficient matrix H . In the early days, NMF was often used to integrate multi-omics data to reveal the hidden patterns and biological meanings in multi-omics data. Zhang et al. proposed the joint NMF algorithm (JNMF) to extract common miRNA-gene-methylation data modules [6]. Considering the correlation between different multimodal data, Zhang et al. proposed the joint sparse network regularization constraint NMF (JSNMNMF) [7]. They added the adjacency matrix among the data as prior knowledge to the JNMF model, making the results more biologically interpretable. In recent years, NMF and its various optimization algorithms have begun to play a role in the field of image genetics. Deng et al. use JSNMNMF to identify ceRNA co-module. They recently added orthogonal constraints to the JSNMNMF and applied it to soft tissue sarcoma research [8]. However, the study only used one type of imaging data and one type of genetic data and did not combine more omics data to conduct in-depth research on the complex mechanisms of the disease. Wang et al. added group-level structure information in the data set to JNMF, proposed group sparse joint non-negative matrix factorization (GSJNMF), and applied it to schizophrenia [9]. On this basis, Peng et al. proposed group sparse of joint non-negative matrix factorization on orthogonal subspace (GJNMFO), which simultaneously performs orthogonal sparse constraint decomposition in the matrix and projects multi-mode data into a low-dimensional orthogonal space [10]. The above two algorithms add group sparse information to NMF to identify hidden dependent structures among different data. However, they ignored the brain connection information. In addition, in our previous research, the Multi-Constrained Joint Non-Negative Matrix Factorization (MCJNMF) was proposed to integrate FDG-PET imaging and DNA methylation data functions. It successfully verified the recognition of soft tissue sarcoma (STS) lung metastasis module [11]. Then, to further explore the relationship between histopathological imaging and genomic data, we proposed the Multi-Dimensional Constrained Joint Non-Negative Matrix Factorization (MDJNMF) [12]. This method effectively discovered the biological function modules related to sarcoma or lung metastasis and reveals the significant correlation between imaging features and genetic variation features.

As far as we know, most imaging genetics research focuses on the correlation analysis between brain imaging and SNP. However, gene expression data can be used to identify disease-causing genes significantly related to AD, which is of great significance for studying the complex mechanisms of diseases. In addition, brain connectivity information can fully reflect the degree of

degeneration and randomization of the brain structure of AD patients. It can explore highly relevant weighting networks of ROIs, SNPs and genes that can detect sets of risk SNPs and genes positively correlated with AD. Therefore, brain and genetic connectivity information are taken into consideration as the prior knowledge in our study. It is of great significance to detect the relationship between gray matter volume changes and genetic information variations and explore a set of highly correlated ROIs/SNPs/Genes closely related to AD/MCI. To integrate sMRI, SNPs, and gene expression data in the ADNI database (<http://ADNI.loni.usc.edu/>), we propose an improved joint negative matrix factorization (NMF) method, named Joint-Connectivity-Based sparse nonnegative matrix factorization (JCB-SNMF). Considering the correlation and the connectivity information in the biological network (including genetics and imaging data), we use the adjacency matrix and GraphNet regularizer [13] as the network regularization constraints to improve the accuracy and noise immunity of the algorithm, respectively. It is a modified version of elastic net regularization, which can effectively integrate physiological constraints such as connectivity. Its stability and anti-interference have been proved in the JCB-SCCA algorithm [14]. GraphNet regularizer is used to implement connectivity of brain and genetic to the algorithm. It encourages related nodes (ROIs, SNPs or genes) to be more similar. By adding it to the coefficient matrix H , we get more relevant features in the significant modules. The results show that with brain connectivity information, TOP 10 pairs of SNP-ROI and gene-ROI contain multiple identical ROIs, SNPs and genes. Through the biological analysis of the significant genes, including Gene Ontology (GO) enrichment analysis and PPI network construction on the selected genes, we found that the selected genes participate in the biological process, and ROI is closely related to neurodegenerative diseases such as AD. We performed the Receiver Operating Characteristic (ROC) analysis on the most interacting genes. We found the potential genes for diagnosis of AD and MCI, including splicing factor 3b subunit 1 (SF3B1), ribosomal protein S20 (RPS20), RNA binding motif protein 14 (RBM14). Besides, most of the genes in which the SNPs in the significant module are closely related to AD. We draw a heat map of the selected SNP-ROI pairs and gene-ROI pairs and found significant relationship pairs, which may be biomarkers of AD and MCI.

2. Method

2.1 Joint Nonnegative Matrix Factorization (JNMF)

NMF is a traditional dimensionality reduction method, and its general model is as follows.

$$\min_{W,H} \|X - WH\|_F^2 \quad s. t. \quad W, H \geq 0 \quad (1)$$

Where $X \in R^{n \times p}$ represents the original feature matrix, which can be decomposed into W and H by NMF. $W \in R^{n \times k}$ is called the basis matrix. $H \in R^{k \times p}$ is called the coefficient matrix. In addition, n is the number of samples, and p is the features of samples. NMF can sufficiently reduce the dimensionality of a single data. However, it cannot be executed on multi-modal data at the same time. Therefore, JNMF is proposed to solve the problem [6]. Its model is as follows.

$$\min_{W,H} \left(\sum_{l=1}^3 \|H_l - WH_l\|_F^2 \right) \quad s. t. \quad W_l, H_l \geq 0 \quad (2)$$

Among them, $X_i \in R^{n \times p_i}$ ($i=1, 2, \dots$) represents the original matrix of different modes. $W_i \in$

$R^{n \times k}$ represents the basis matrix obtained by decomposition. $\mathbf{H}_I \in R^{k \times p_i}$ represents the coefficient matrix of each original matrix obtained by decomposition.

2.2 Joint Sparse Network-Regularized Multiple Nonnegative Matrix Factorization (JSNMNMF)

Considering that the coefficient matrix obtained by JNMF has strong independence and n is often much smaller than p in practical applications, [7] proposed the JSNMNMF. JSNMNMF adds prior knowledge to the objective function to improve the biological relevance of the results and can also improve the efficiency of the module by reducing the larger search space. In the paper, to improve the weak connection between imaging and genetics, Assume $\mathbf{A}_1$ is the MRI-SNP interaction adjacency matrix, $\mathbf{A}_2$ is the MRI-Gene interaction adjacency matrix, and $\mathbf{A}_3$ is the SNP-Gene interaction adjacency matrix. In addition, to sparse the data to discover the key features of the data, JSNMNMF uses the method of [15] to control the sparsity of $\mathbf{W}$ and $\mathbf{H}$. Therefore, its objective function is as follows.

$$\begin{aligned} \Gamma(\mathbf{W}, \mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3) = & \sum_{I=1}^3 \|\mathbf{X}_I - \mathbf{W}\mathbf{H}_I\|_F^2 - \lambda_1 \text{Tr}(\mathbf{H}_1 \mathbf{A}_1 \mathbf{H}_2^T) - \lambda_2 \text{Tr}(\mathbf{H}_1 \mathbf{A}_2 \mathbf{H}_3^T) \\ & - \lambda_3 \text{Tr}(\mathbf{H}_2 \mathbf{A}_3 \mathbf{H}_3^T) + \gamma_1 \|\mathbf{W}\|_F^2 + \gamma_2 \left(\sum_{I=1}^3 \|\mathbf{H}_I\|_F^2 \right) \end{aligned} \quad (3)$$

Where the parameter λ_1, λ_2 , and λ_3 are the weights for the adjacency constraints. γ_1 is used to limits the growth of $\mathbf{W}$ and γ_2 is used to constrain $\mathbf{H}$.

2.3 Joint-Connectivity-Based Sparse Nonnegative Matrix Factorization (JCB-SNMF)

To encourage the similarity of related elements of the norm vector, the connectivity-based penalty term [14] is introduced in the JSNMNMF algorithm. Specifically, suppose the connectivity between the i -th node and the j -th node (that is, the brain region or genetic data) is high. In that case, it will force the corresponding elements of the norm vector to be similar [13]. Therefore, we add the brain connectivity information and the weighted SNP correlation network to capture the genetic network structure [16] as a prior matrix to add to the algorithm, aiming to improve the biological significance of the extracted features.

$$\begin{aligned} P(\mathbf{H}_1) &= \sum_{p,q} L_{H_1(p,q)} (\mathbf{H}_{1p} - \mathbf{H}_{1q}) \\ P(\mathbf{H}_2) &= \sum_{p,q} L_{H_2(p,q)} (\mathbf{H}_{2p} - \mathbf{H}_{2q}) \\ P(\mathbf{H}_3) &= \sum_{p,q} L_{H_3(p,q)} (\mathbf{H}_{3p} - \mathbf{H}_{3q}) \end{aligned} \quad (4)$$

Here, $\mathbf{L}_{h_1}$, $\mathbf{L}_{h_2}$ and $\mathbf{L}_{h_3}$ represent the Laplacian matrix of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, respectively.

We regard the Laplacian matrix of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ as a new fusion penalty of JSNMNMF algorithm, and the specific formula is as follows:

$$\begin{aligned} P(\mathbf{H}_1) &= \text{Tr}(\mathbf{H}_1^T \mathbf{B}_1 \mathbf{H}_1) \\ P(\mathbf{H}_2) &= \text{Tr}(\mathbf{H}_2^T \mathbf{B}_2 \mathbf{H}_2) \\ P(\mathbf{H}_3) &= \text{Tr}(\mathbf{H}_3^T \mathbf{B}_3 \mathbf{H}_3) \end{aligned} \quad (5)$$

B_1 , B_2 and B_3 represent the Laplacian matrices of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, respectively.

2.4 The Efficient Optimization Algorithm

Now we can write the proposed method with penalties explicitly exhibited.

$$\begin{aligned} \Gamma(W, H_1, H_2, H_3) = & \sum_{l=1}^3 [T_r(X_l X_l^T) - 2T_r(X_l H_l^T W^T) + T_r(W H_l H_l^T W^T)] - \lambda_1 T_r(H_1 A_1 H_1^T) \\ & - \lambda_2 T_r(H_1 A_2 H_1^T) - \lambda_3 T_r(H_2 A_3 H_2^T) + \beta_1 T_r(H_1^T B_1 H_1) + \beta_2 T_r(H_2^T B_2 H_2) \\ & + \beta_3 T_r(H_3^T B_3 H_3) + \gamma_1 \|W\|_F^2 + \gamma_2 \sum_{l=1}^3 e_{1 \times k} H_l H_l^T e_{1 \times k}^T \end{aligned} \quad (6)$$

Let φ_{ij} and ϕ_{ij}^l be constrained Lagrange multipliers of $[W_{ij} \geq 0 \ \& \ (H_l)_{ij} \geq 0]$.

$$L(W, H_1, H_2, H_3) = \Gamma(W, H_1, H_2, H_3) + T_r(\Psi W^T) + \sum_{l=1}^3 T_r(\Phi H_l^T) \quad \Psi = [\varphi_{ij}], \quad \Phi_l = [\phi_{ij}^l] \quad (7)$$

The partial derivatives of L with respect to W and H_l are:

$$\begin{aligned} \frac{\partial L}{\partial W} &= \sum_{l=1}^3 [-2X_l H_l^T + 2W H_l H_l^T] + 2\gamma_1 W \\ \frac{\partial L}{\partial H_1} &= -2W^T X_1 + 2W^T W H_1 + \lambda_1 H_2 A_1^T + \lambda_2 H_3 A_2^T + 2\gamma_2 e_{k \times k} H_1 + 2\beta_1 H_1 C_1 \\ \frac{\partial L}{\partial H_2} &= -2W^T X_2 + 2W^T W H_2 + \lambda_1 H_1 A_1 + \lambda_3 H_3 A_3^T + 2\gamma_2 e_{k \times k} H_2 + 2\beta_2 H_2 C_2 \\ \frac{\partial L}{\partial H_3} &= -2W^T X_3 + 2W^T W H_3 + \lambda_2 H_1 A_2 + \lambda_3 H_2 A_3 + 2\gamma_2 e_{k \times k} H_3 + 2\beta_3 H_3 C_3 \end{aligned} \quad (8)$$

Based on Karush-Kuhn-Tucher (KKT) conditions, $\Psi_{ij} W_{ij} = 0 \ \& \ \Phi_{ij}^l (H_l)_{ij} = 0$. We can obtain the equations of W_{ij} and $(H_l)_{ij}$:

$$\begin{aligned} (W)_{pq} &= (W)_{pq} \frac{(X_1 H_1^T + X_2 H_2^T + X_3 H_3^T)_{pq}}{(W H_1 H_1^T + W H_2 H_2^T + W H_3 H_3^T + \gamma_1 W)_{pq}} \\ (H_1)_{pq} &= (H_1)_{pq} \frac{\left(W^T X_1 + \frac{\lambda_1}{2} H_2 A_1^T + \frac{\lambda_2}{2} H_3 A_2^T\right)_{pq}}{(W^T W H_1 + \beta_1 H_1 B_1 + \gamma_2 e_{k \times k} H_1)_{pq}} \\ (H_2)_{pq} &= (H_2)_{pq} \frac{\left(W^T X_2 + \frac{\lambda_1}{2} H_1 A_1 + \frac{\lambda_3}{2} H_3 A_3^T\right)_{pq}}{(W^T W H_2 + \beta_2 H_2 B_2 + \gamma_2 e_{k \times k} H_2)_{pq}} \\ (H_3)_{pq} &= (H_3)_{pq} \frac{\left(W^T X_3 + \frac{\lambda_2}{2} H_1 B_2 + \frac{\lambda_3}{2} H_2 A_3\right)_{pq}}{(W^T W H_3 + \beta_3 H_3 B_3 + \gamma_3 e_{k \times k} H_3)_{pq}} \end{aligned} \quad (9)$$

2.5 Selection of module elements

Through the iterative update of the above algorithm, we finally decompose the feature matrix $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ of MRI, SNP, and gene into base matrix W and coefficient matrix H_1 , H_2 and H_3 . In order to find the weight value corresponding to the distinctive feature of each row of W , we use the z-score to extract the coefficient of each row of the H_l matrix. It is defined as follows.

$$Z_{ij} = \frac{(h_{ij} - u_i)}{\sigma_i} \quad (9)$$

Where h_{ij} represents the element in H_l , u_i represents the average value of feature j in H_l , and σ_i

represents the standard deviation. Next, in order to determine the module membership, we set a threshold T . If its z-score is greater than the set threshold T , it is considered eligible to be assigned to the module.

2.6 Estimate the significance of the co-expression module

In this part, we use the method of [5] to evaluate the significance of each module. Specifically, assume that $A^Q = [a_1, a_2, \dots, a_{l_1}]$, $B^Q = [b_1, b_2, \dots, b_{l_2}]$ and $C^Q = [c_1, c_2, \dots, c_{l_3}]$ where a_g , b_t , and c_e are column vectors selected from X_1 , X_2 , and X_3 . Then, the mean correlations ρ^* among the three types of datasets in a module can be expressed as follows.

$$\rho^* = \frac{1}{3} \left(\frac{1}{l_1 l_2} \sum_{g=1}^{l_1} \sum_{t=1}^{l_2} (\rho(a_g, b_t))^2 + \frac{1}{l_1 l_3} \sum_{g=1}^{l_1} \sum_{e=1}^{l_3} (\rho(a_g, c_e))^2 + \frac{1}{l_2 l_3} \sum_{t=1}^{l_2} \sum_{e=1}^{l_3} (\rho(b_t, c_e))^2 \right) \quad (10)$$

For a given matrix C^Q , we randomly change the order of the row vectors of the matrices A^Q and B^Q in Q -th module, and repeat this process Δ times. For each permutation, ρ_θ^* is the new mean correlation coefficient calculated by (10) after the row of permutation matrix A^Q and B^Q . The significance of the test statistic can be estimated as follows.

$$p - \text{value} = \frac{|\{\theta | \rho_\theta^* \geq \rho^*, \theta = 1, 2, \dots, \Delta\}|}{\Delta} \quad (11)$$

Where $|\cdot|$ denotes the number of times $\rho_\theta^* \geq \rho^*$. If the P-value is less than 0.05, we consider the module is significant.

3. Results

3.1 Data source and preprocessing

In this section, we evaluate the effectiveness of the proposed method on the ADNI database. In this database, we examine and select candidate SNPs to predict the sMRI phenotype reaction. As can be seen from Table 1, there were 180 non-Hispanic Caucasian participants with imaging and genotyping data, including 21 healthy controls (HC), 147 mild cognitive impairment (MCI), and 12 AD Participants.

Table 1 Characteristics of the subjects			
Groups	AD	MCI	HC
Number	12	147	21
Gender(M/F)	6/6	78/49	10/11
Age(mean $\pm$ std)	79.5 $\pm$ 10.22	71.56 $\pm$ 7.51	72.32 $\pm$ 7.76

Note: HC=Healthy Control, MCI=Mild Cognitive Impairment, AD=Alzheimer's disease.

The original sMRI imaging was downloaded from ADNI1 in the experiment, which used DiffusionKit [17] software to achieve head movement correction. They registered it to the Montreal Neurological Institute (MNI) standard space. Next, the segmentation of sMRI imaging was implemented by using MATLAB software CAT toolkit in the SPM software package [18]. Specifically, voxel-based morphometry (VBM) provides voxel estimation of the local number or volume of specific tissue compartments. By scaling the volume change due to spatial registration to adjust the segmentation, the volume of gray matter tissue was calculated in the region of interest (ROI) as a feature. After screening, in the end, 140 ROIs were retained.

The genotypes of 180 subjects in this study came from the ADNI1 database. The genome-

wide SNP sites of samples were screened through the following steps. All SNPs were genotyped by the human 610 Quad BeadChip in the study. We used the genetic analysis tool PLINK [19] to screen genotype data, using the following exclusion criteria: rare SNPs (minor allele frequency (MAF) < 0.05), violations of Hardy-Weinberg Equilibrium (HWE $p < 10^{-6}$), poor call rate ($< 90\%$) per subject and per SNP marker, gender check, and sibling pair identification. This resulted in the final dataset spanning 5947 SNP loci.

We use the limma package [20] to screen for genes with significant differential expression. Those genes were removed when the p-value was higher than 0.01, 1477 genes were obtained.

For the feature matrices of the three original data obtained by the above processing, we use the L2 norm to normalize the data to ensure the non-negativity of the input data. Specifically, since NMF is an unsupervised clustering method, we only put three feature matrices composed of samples from the AD group and the MCI group into the algorithm. The samples of the HC group were used as controls during the preprocessing of SNP and gene expression data. Then we spliced the feature matrices of the AD group and MCI group samples together. Finally, we obtain three characteristic matrices $\mathbf{X}_1 \in R^{159 \times 140}$, $\mathbf{X}_2 \in R^{159 \times 5947}$ and $\mathbf{X}_3 \in R^{159 \times 1477}$ corresponding to sMRI, SNP and gene expression data.

3.2 Parameter Selection

In this part, we use the proposed algorithm to do real data experiments. We apply the proposed algorithm to SNPs, sMRI and gene expression data. According to the formulas (6) in the Method, we need to determine the optimal λ_i , β_i ($i=1,2,3$), γ_1 , and γ_2 for the proposed algorithm. In the experiment, we randomly initialize a set of non-negative $\mathbf{W}$ and $\mathbf{H}_I$ ($I=1, 2, 3$). First, to prove the convergence of the proposed algorithm, we randomly select three sets of parameters and iterate 100 times. Then we plot the objective function values obtained during the iteration process in Fig. 1. It can be seen from Fig. 1 that the objective function value of the proposed algorithm is greatly affected by parameter adjustment. However, a blind grid search of parameters is highly time-consuming. Thus, we fix the value of K to 10 and tune λ_i , β_i ($i=1,2,3$), γ_1 , and γ_2 from the following finite set: $[0.0001, 0.001, 0.01, 0.1, 1, 10]$. Fig. 2 is the reconstruction error obtained by sequentially substituting 1296 sets of regularization parameters into the algorithm. We use the set of regularization parameters corresponding to the minimum reconstruction error for further analysis.

Fig.1 The value of the objective function recorded with 100 iterations under the condition of three sets of randomly selected parameters

Fig. 2 Reconstruction errors obtained under different combinations

As shown in Fig. 2, when taking the 866th group of regularization parameters, the reconstruction error is the smallest. Parameters in group 866 are $\lambda_i = 0.001$, $\beta_i = 0.0001$ ($i=1,2,3$), $\gamma_1 = 0.1$, and $\gamma_2 = 0.0001$. Then, we fix λ_i , β_i , γ_1 , and γ_2 and use the same method to determine K . The algorithm based on NMF requires $K \ll n$. We set the upper limit of K to 50. The proposed model reconstructed data with different dimensionality, as shown in Fig. 3. When K increases, the reconstruction error will gradually decrease.

However, when K is less than 6, we find that some common modules no longer contain any elements. Therefore, the K value is chosen to be between six and fifty. When the K value is too large, it will cause changes of the overlap rate among modules, which is not conducive to subsequent analysis. Specifically, different K values will affect the Pearson correlation coefficients between the original and decomposed matrices and the element overlap ratio among the modules. Therefore, we compared the two indicators under different K values.

Specifically, we used the overlap rate calculation method consistent with previous research [12]. The overlap rate is defined as $O = \frac{\text{length}(\text{insert}(R_i, R_j))}{\min(\text{length}(R_i), \text{length}(R_j))}$, where R_i and R_j represent the feature vectors of the i -th and j -th results respectively, $\text{length}()$ represents the number of feature vectors, and $\text{insert}()$ represents the intersection of the i -th result and the j -th result. As shown in Fig.4, (A)-(C) respectively show the overlap rate of ROIs, SNPs, and genes modules under different K values. As the K value increases, the overlap rate of members in the ROI and gene modules gradually decreases. However, the overlap rate of members in the SNP modules is gradually

increasing. (D)-(F) respectively show the Pearson correlation coefficients between X_i and W^*H_i ($i=1,2,3,\dots$) under different K values. As the value of K increases, the Pearson correlation coefficients between X_i and X_2 and their respective reconstruction matrices gradually increase. However, the Pearson correlation coefficient between X_3 and its reconstruction matrix gradually decreases. When the number of modules is about 30, it can be considered as a balance point based on Fig. 4 (A)-(F). Based on the above considerations, we set K to 30.

Fig. 3 Reconstruction errors obtained under different K value

Fig. 4 (A)-(C) respectively show the overlap rate of ROIs, SNPs, and genes modules under different K values. (D)-(F) respectively show the Pearson correlation coefficients between X_i and W^*H_i ($i=1,2,3,\dots$) under different K values.

To prevent the objective function from falling into a local minimum, we repeat the whole procedure 100 times with different initialization values. When the objective function takes the minimum value, the corresponding W , H_1 , H_2 , and H_3 are used for further analysis. We have drawn the objective values, as shown in Fig. 5. Finally, we selected the 18th set of initialization values. In addition, we added two indicators previously selected for K to prove the algorithm's multiple

experiment replication rate. As shown in Fig.6, as the initial value increases, the overlap ratio of the three module members and the Pearson correlation coefficients of the three matrices and their respective reconstruction matrices are relatively stable. The average overlap ratios of the three module members are 0.7860, 0.6776 and 0.7443, respectively. Thus, it shows that the proposed algorithm is robust and reproducible.

Fig. 5 The value of objective function with different initialization value

Fig. 6 (A)-(C) show the overlap rate of ROIs, SNPs, and genes modules under different initialization values, respectively. (D)-(F) show the Pearson correlation coefficients of X_i and W^*H_i ($i=1,2,3,\dots$) under different initialization values, respectively.

3.3 Results on ADNI database

We conducted simulation experiments on the ADNI database preprocessed above. The Pearson correlation coefficients of the original matrix and the reconstructed matrix of ROI, SNP, GENE are 0.9962, 0.5534, 0.9921, respectively. Thirty co-expression modules were also obtained. In addition, we calculated the computational cost of the three algorithms. We run the experiments of real datasets on a machine with IntelCore i5-8300H CPU with 16 GB RAM and used MATLAB (R2017a) 64-bit for the general implementation. The computational costs of the proposed algorithm, JNMNMF and JNMF are 19.02s, 13.46s, 10.23s, respectively.

In order to verify the correlation analysis capability of the proposed algorithm, we performed KEGG enrichment analysis on the gene expression data in all modules. We extracted the seven most enriched biological process keywords and calculated the number of keywords involved in each module. It can be seen from Fig. 7 that there are 19 modules involving all seven keywords. Toll-like receptors can initiate pro-inflammatory immune responses by activating NF- κ B and other transcription factors that cause the synthesis of pro-inflammatory molecules and play an important role in diseases related to neuroinflammation (AD, Parkinson's disease, etc.) [21]. [22] has shown that immune cells in mice infected with toxoplasmosis promote neuroinflammation through a cytokine network and enhance cognitive impairment in AD mice. Influenza A is also closely related to AD. The accumulation of β -amyloid (β A) can cause aggravation of AD, but it also inhibits the influenza A virus [23]. Therefore, the proposed algorithm can effectively select disease-related modules, and different modules have a certain degree of representation.

Fig.7 Top seven enriched biological process and the number of modules involved in these processes

We used Formula (11) in the Method section to select the five modules with $p < 0.05$, as shown in Table. 2 Among them, Module 6 and Module 10 contain SNPs that exceed 60% of the total number of SNPs are not considered. Draw a Venn diagram for the three data features in other modules, and compare the escape rates of the three data features in different modules. As shown in Fig. 8, the escape rate of module 1 is the smallest among the three types of data, so we choose module 1 for further analysis.

Table 2 Module selection

Module	ROI	SNP	GENE	p-value
6	28	2401	292	6.86E-34
1	29	128	258	7.12E-29
20	28	800	285	6.82E-04
10	27	1940	274	0.0228
17	31	514	307	0.0450

Fig. 8 The overlap of ROIs, SNPs and genes in the three modules A. The overlap of ROIs among three modules B. The overlap of SNPs among three modules C. The overlap of genes among three modules

Table 3 Selected ROIs and genes (from SNPs data) in Module 1

ROIs	Gene ID
Left/Right Exterior Cerebellum	GEMIN6 DHX57 ARHGEF33
Left/Right Cerebral White Matter	CAMKMT SAP130 IL5RA
Left/Right Angular Gyrus	DMD EOMES KLF15 SLC9A9
Right Fusiform Gyrus	PPP2R2C IL1RAPL2 TRMT44
Left/Right Inferior Occipital Gyrus	LOC107986262 LOC105379102
Left/Right Inferior Temporal Gyrus	ADCYAP1R1 GPR141 NOL4
Left/Right Lingual Gyrus	GIMAP7 LINC00824 KCNC2
Left/Right Middle Frontal Gyrus	ASXL3 CNTNAP3 LOC107987084
Left/Right Middle Temporal Gyrus	LOC107987122 NUP214
Left/Right Precuneus	LHFPL6 LINC02846
Left/Right Postcentral Gyrus	LOC101929846 FAM24B CASK
Left/Right Precentral Gyrus	CUZD1 C10orf88B PSMB3
Left/Right Superior Frontal Gyrus	C10orf88 PTH SPON1
Left Supramarginal Gyrus	HSD17B12 PCSK2 CREB3L1
Left/Right Superior Parietal Lobule	AMBRA1 ATG13 LOC107984384
Left Superior Temporal Gyrus	LOC105369520 LOC105370234
	DLGAP1 RGS6 LOC107983974

4 Discussion

4.1 Biological significance

Table 3 lists the ROIs and genes identified by the proposed algorithm from the SNPs loci. As can be seen from Table 3, Cerebral White Matter [24], Angular Gyrus [25], Fusiform Gyrus [26], Inferior Occipital Gyrus [27], Inferior Temporal Gyrus [28], Lingual Gyrus [29], Middle Frontal Gyrus [30], Middle Temporal Gyrus [31], Precuneus [32], Postcentral Gyrus [33], Precentral Gyrus [34], Supramarginal Gyrus [35], Superior Parietal Lobule [36], Superior Temporal Gyrus [37] have been confirmed to be risk brain regions for AD. Also, we identify a total of 47 risk genes from the SNPs data in the first module. Literature [38] [39] found that DHX57 and SPON1 associated with

accelerated cerebellar age significantly affect AD, respectively. In another two imaging genetics studies, EOMES and RGS6 were also identified as risk genes [40][41]. The unregulated expression of PPP2R2C may be related to the onset of AD [42]. The amyloid precursor protein (APP) plays a central role in AD, and CASK is an interactor of the APP intracellular domain (AICD) [43]. [44] has confirmed that in the mouse model of tauopathy, PCSK2 is up-regulated as a differential gene in its hippocampus. CNTNAP5 comes from the contactin-associated protein (Caspr) family and is related to various neurodegenerative diseases [45]. Besides, SLC9A9 is also associated with neuropsychiatric diseases [46].

We found several SNP sites related to senile and neurological diseases in Module 1. Frailty is a complex phenotype of aging. A study has found that the prevalence of haplotypes of risk alleles on rs1324192 is significantly higher than that of non-vulnerable older adults [47]. The intron SNP rs3802890 is associated with female autism characteristics [48]. [49] found that rs981975 is the highest SNP correlated with changes in the anti-schizophrenia dose. [50] predicted that rs12185438 is highly related to Parkinson's syndrome through machine learning.

Fig. 9 The result of GO enrichment analysis for the genes selected in Module 1. The horizontal axis is the number of genes in the pathway, the vertical axis is the pathway list, and the red to blue color indicates that the p-value ranges from 0 to 10.

Fig.10 PPI network constructed with genes selected in Module 1

258 genes are selected from the gene expression data set. As shown in Fig. 9, we performed GO enrichment analysis on these genes. [51] has confirmed that dysfunction of the nucleoskeleton is causal for Alzheimer's disease-related neurodegeneration. [52] has confirmed that activity of the poly(A) binding protein MSUT2 determines susceptibility to pathological tau in the mammalian brain. Most of the other enriched biological processes are also closely related to AD. We also constructed a protein-protein interaction network (PPI) for the genes selected in Module 1. As can be seen in Fig. 10, we have retained 177 genes with interaction relationships. We selected ten genes with the largest area (strongest interaction), namely SF3B1, BPTF, TLR4, MTOR, LSM3, SRRM1, PCBP1, HNRNPA3, RPS20, and RBM14. Among them, [53] has confirmed that in the initial stage, the activation of TLR4 has an effective clearance effect on amyloid β ($A\beta$). However, the long-term activation state will cause $A\beta$ to be deposited in the brain. [54] points out that AD can be prevented and treated by regulating the MTOR signaling pathway. [55] confirmed that LSM3 is the main pathogenic gene of AD and the core gene in the module network closely related to MCI and AD. The biological functions of HNRNPA3 products are related to inflammation and neurodegenerative diseases [56]. Those significant genes were analyzed by Receiver Operating Characteristic (ROC) curve using IBM SPSS Statistics 22. We show the area under the Curve (AUC) values greater than 0.5 and p-value smaller than 0.05 in Fig. 11 and Table 4. In Fig. 10, we show the ROC curves of three significant genes. In Table 4, we show the detailed information of the ROC curves of the three genes. There are four genes with the area under the Curve (AUC) values greater than 0.6 and significant values below 0.01. The highest AUC was found for RBM14 (AUC: 0.691, 95% CI: 0.577-0.805, P = 0.007), followed by SF3B1 (AUC: 0.656, 95% CI: 0.525-0.788, P = 0.026), and RPS20 (AUC: 0.641, 95% CI: 0.504-0.777, P = 0.045). Therefore, the proposed algorithm verified several AD and MCI risk genes in Module 1 and found three potential genes related to AD and MCI.

Fig.11 ROC analysis of significant genes. A. ROC curve of SF3B1. B. ROC curve of RPS20. C. ROC curve of RBM14.

Table 4 ROC curve information of four significant genes

Gene ID	AUC	P value	95%CI
SF3B1	0.656	0.026	0.525-0.788
RPS20	0.641	0.045	0.504-0.777
RBM14	0.691	0.007	0.577-0.805

According to the SNPs, genes, and brain ROIs selected in Module 1, Fig. 12 shows the pairwise correlation heat map of brain ROI-SNP/gene pairs. As expected, most ROI-SNP/gene pairs are

strong. To find significantly stronger SNP/gene-ROI pairs, we show the top 10 pairs-with $p < 0.01$ in Module 1 from Table 5. Although no SNP in the SNP-ROI pair has been reported yet, since most of the ROI has been confirmed to be the risk brain area of AD/MCI, the obtained SNP-ROI pair still has a specific reference value. Combining our conclusions and reports in the literature [24], the variation of rs4472239 may play an important role in the abnormality of Cerebral White Matter, thereby promoting the development of MCI and AD. rs11918049 is significantly associated with multiple frontal regions and the Central Anterior Gyrus. The genes in the TOP 10 pairs of Gene-ROI are concentrated in KLHL8, ZC3H11A and OSGEPL1. As part of the topology described in the literature [57], ZC3H11A can cause immune system disorders related to aging. [58] have shown that there is no significant change in the expression of KEOPS complex (including OSGEPL1) in the brains of patients with depression and schizophrenia, but whether the KEOPS complex has functional changes in neuropsychiatric diseases has not been confirmed. The differential expression of ZC3H11A and functional changes of OSGEPL1 in HC and AD/MCI may affect the gray matter volume of the Superior Parietal Lobule. There is no research to confirm the effect of KLHL8 on neurodegenerative diseases. Therefore, the conclusion that it impacts multiple ROIs in our research remains to be confirmed.

Fig. 12 The heat map for SNP-ROI and Gene-ROI. A. The heat map of SNP-ROI associated with

the selected markers in Module 1. B. The heat map of Gene-ROI associated with the selected markers in Module 1.

Table 5 The top ten pairs of SNP-ROI and Gene-ROI with $p < 0.01$ in module 1, respectively.

SNP-ROI	P Value	GENE-ROI	P Value
rs4472239-lCbrWM	2.60E-16	KLHL8-rSupParLo	1.73E-13
rs11918049-lPrcGy	9.73E-14	KLHL8-rMidTemGy	8.78E-13
rs11918049-rSupFroGy	2.07E-13	ZC3H11A- rSupParLo	2.82E-12
rs4472239-rCbrWM	3.5E-13	KLHL8-lMidTemGy	5.23E-12
rs1077971- rExtCbe	9.37E-13	KLHL8- lAngGy	2.39E-11
rs11918049-rMidFroGy	1.94E-12	KLHL8- lSupMarGy	2.82E-10
rs2446856- lCbrWM	1.32E-11	OSGEPL1- lSupParLo	4.29E-10
rs1324192- rCbrWM	1.01E-10	KLHL8- lSupTemGy	8.55E-10
rs11918049-rPrcGy	2.53E-10	OSGEPL1- rExtCbe	8.76E-10
rs11918049-lMidFroGy	4.35E-10	KLHL8-lInfTemGy	1.3E-9

Note: lCbrWM =Left Cerebral White Matter; lPrcGy= Left Precentral Gyrus; rSupFroGy= Right Superior Frontal Gyrus; rCbrWM= Right Cerebral White Matter; rExtCbe=Right Exterior Cerebellum; rMidFroGy =Right Middle Frontal Gyrus; lPrcGy=Right Precentral Gyrus; lMidFroGy=Left Middle Frontal Gyrus; rSupParLo=Right Superior Parietal Lobule; rMidTemGy=Right Middle Temporal Gyrus; lMidTemGy=Left Middle Temporal Gyrus; lAngGy= Left Angular Gyrus; lSupMarGy=Left Supramarginal Gyrus; lSupParLo=Left Superior Parietal Lobule; lSupTemGy=Left Superior Temporal Gyrus; lInfTemGy=Left Inferior Temporal Gyrus

4.2 Comparison of anti-noise performance with other algorithms

This research used the Laplacian matrix as a new penalty item to process the image and genetic data. It can improve stability and anti-noise performance. We conduct experiments on two data sets with larger and smaller sample sizes and features, respectively. We use n to represent the number of samples, $p1$ to represent sMRI features, $q1$ to represent SNP features, and $q2$ to represent gene features. In the large set, we set $n=300$, $p1=2000$, $q1=2500$, $q2=1000$, respectively. In the small set, we set $n=100$, $p1=650$, $q1=350$, $q2=600$, respectively. In addition, we set $K=10$. The elements in the base matrix W and the coefficient matrix H_1 , H_2 , and H_3 are all random integers. W , H_1 , H_2 , and H_3 are generated by the following equation [5].

$$\alpha[n] = \{\beta_i | \beta_i = \alpha + l\eta_i (i = 1, 2, \dots, n)\}$$

α is a matrix made of random integers from a uniform distribution $U(1, 10)$. Then, we use η_i to represent the Gauss noise and l to represent the noise level.

It can be seen from Fig.13 that in the process of increasing noise in the two data sets, the reconstruction error and objective function value of the proposed algorithm are both smaller than JNMF and JSNMNMF. In addition, we calculated the computational cost of the three algorithms. We run the experiments of synthetic datasets on a machine with IntelCore i5-8300H CPU with 16 GB RAM and used MATLAB (R2017a) 64-bit for the general implementation. In the small set, the computational costs of the proposed algorithm, JNMNMF and JNMF are 7.49s, 7.33s, 7.28s, respectively. In the large set, the computational costs of the proposed algorithm, JNMNMF and JNMF are 16.53s, 15.14s, 12.49s, respectively.

Fig. 13 The reconstruction error and objective function value of these three algorithms under different noise levels are compared. A and B are used for small data sets. C and D are used for large data sets.

5 Conclusion

NMF is a robust dimensionality reduction analysis algorithm that can integrate multiple omics data. Therefore, introducing it into image genetics can effectively integrate the macroscopic and microscopic characteristics of the disease and then mine the biomarkers closely related to the disease from the significant co-expression modules of the characteristics. In the paper, we proposed JCB-SNMF. This method considers the respective connectivity information of the brain and genes. Furthermore, it adds the Laplacian matrix as prior knowledge to the JSNMNMF algorithm, which improves the algorithm's anti-noise performance and biological interpretability.

The proposed algorithm uses sMRI, SNP, and gene expression data in the ADNI dataset. Simulation results show that the noise resistance performance of the proposed algorithm is better than JSNMNMF and JNMF. Experimental results on real data show that the proposed algorithm can identify and predict risk ROI, risk SNPs, and risk genes closely related to AD and MCI. Not only that, but we also found some significant SNP/Gene-ROI pairs. For example, rs11918049 is related to multiple frontal regions and the Central Anterior Gyrus. In addition, ZC3H11A and OSGEPL1 are closely related to changes in gray matter volume in multiple brain regions. In the future, we will integrate multi-modal imaging (PET, CT, etc.) and genetic data (DNA methylation, etc.) to discover more complex biological mechanisms closely related to diseases.

ETHICAL STATEMENT

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants or animals performed by any of the authors; therefore, the ethical approval and consent to participate are not applicable.

Consent for publication

The authors have consented to publication of this article.

Availability of data and materials

sMRI, SNPs, and gene expression data of patients with Alzheimer disease and controls were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu/>)

Competing interests

The authors declare that they have no competing interests.

Funding

The Natural Science Foundation of Shanghai (No. 18ZR1417200) and the National Natural Science Foundation of China (No.61803257).

Authors' contributions

Conception and design of the research: Kai Wei and Wei Kong. Acquisition, analysis, and interpretation of data: Kai Wei, Shuaiqun Wang, and Wei Kong. Statistical analysis: Kai Wei. Drafting the manuscript: Kai Wei. Manuscript revision for important intellectual content: Wei Kong. All authors have read and approved the manuscript.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Shanghai (No. 18ZR1417200) and National Natural Science Foundation of China (No. 61803257).

References

- [1] Parkhomenko, Elena et al. "Sparse canonical correlation analysis with application to genomic data integration." *Statistical applications in genetics and molecular biology* vol. 8 (2009): Article 1. doi:10.2202/1544-6115.1406
- [2] Du, Lei et al. "Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach." *Medical image analysis* vol. 61 (2020): 101656. doi:10.1016/j.media.2020.101656
- [3] Yan, Jingwen et al. "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm." *Bioinformatics (Oxford, England)* vol. 30,17 (2014): i564-71. doi:10.1093/bioinformatics/btu465
- [4] Du, Lei et al. "Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method." *Bioinformatics (Oxford, England)* vol. 32,10 (2016): 1544-51. doi:10.1093/bioinformatics/btw033
- [5] Peng, Peng et al. "Group Sparse Joint Non-negative Matrix Factorization on Orthogonal Subspace for Multi-modal Imaging Genetics Data Analysis." *IEEE/ACM transactions on computational biology and bioinformatics*, vol. PP 10.1109/TCBB.2020.2999397. 2 Jun. 2020, doi:10.1109/TCBB.2020.2999397

- [6] Zhang, Shihua et al. "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data." *Nucleic acids research* vol. 40,19 (2012): 9379-91. doi:10.1093/nar/gks725
- [7] Zhang, Shihua et al. "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules." *Bioinformatics* (Oxford, England) vol. 27,13 (2011): i401-9. doi:10.1093/bioinformatics/btr206
- [8] Deng, Jin et al. "Prior Knowledge Driven Joint NMF Algorithm for ceRNA Co-Module Identification." *International journal of biological sciences* vol. 14,13 1822-1833. 19 Oct. 2018, doi:10.7150/ijbs.27555
- [9] Deng, Jin et al. "Multi-Constrained Joint Non-Negative Matrix Factorization With Application to Imaging Genomic Study of Lung Metastasis in Soft Tissue Sarcomas." *IEEE transactions on bio-medical engineering* vol. 67,7 (2020): 2110-2118. doi:10.1109/TBME.2019.2954989
- [10] Wang, Min et al. "Integration of Imaging (epi)Genomics Data for the Study of Schizophrenia Using Group Sparse Joint Nonnegative Matrix Factorization." *IEEE/ACM transactions on computational biology and bioinformatics* vol. 17,5 (2020): 1671-1681. doi:10.1109/TCBB.2019.2899568
- [11] J. Deng, W. Zeng, W. Kong, Y. Shi, X. Mou and J. Guo, "Multi-Constrained Joint Non-Negative Matrix Factorization With Application to Imaging Genomic Study of Lung Metastasis in Soft Tissue Sarcomas," in *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 7, pp. 2110-2118, July 2020, doi: 10.1109/TBME.2019.2954989.
- [12] Jin Deng, Weiming Zeng, Sizhe Luo, Wei Kong, Yuhu Shi, Ying Li, Hua Zhang, Integrating Multiple Genomic Imaging Data for the study of Lung Metastasis in Sarcomas using Multi-Dimensional Constrained Joint Non-Negative Matrix Factorization, *Information Sciences*, 2021, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2021.06.058>.
- [13] Grosenick, Logan et al. "Interpretable whole-brain prediction analysis with GraphNet." *NeuroImage* vol. 72 (2013): 304-21. doi:10.1016/j.neuroimage.2012.12.062
- [14] Kim, Mansu et al. "Joint-Connectivity-Based Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Parkinson's Disease." *IEEE transactions on medical imaging* vol. 39,1 (2020): 23-34. doi:10.1109/TMI.2019.2918839
- [15] Kim, Hyunsoo, and Haesun Park. "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis." *Bioinformatics* (Oxford, England) vol. 23,12 (2007): 1495-502. doi:10.1093/bioinformatics/btm134
- [16] Levine, Morgan E et al. "A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores." *Methods in molecular biology* (Clifton, N.J.) vol. 1613 (2017): 277-290. doi:10.1007/978-1-4939-7027-8_10
- [17] Gorski, Jochen , F. Pfeuffer , and K. Klamroth . "Biconvex sets and optimization with biconvex functions: a survey and extensions." *Mathematical Methods of Operations Research* 66.3(2007):373-407.
- [18] Saykin, Andrew J et al. "Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans." *Alzheimer's & dementia : the journal of the Alzheimer's Association* vol. 6,3 (2010): 265-73. doi:10.1016/j.jalz.2010.03.013
- [19] Purcell, Shaun et al. "PLINK: a tool set for whole-genome association and population-based

- linkage analyses.” *American journal of human genetics* vol. 81,3 (2007): 559-75. doi:10.1086/519795
- [20] Ritchie, Matthew E et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic acids research* vol. 43,7 (2015): e47. doi:10.1093/nar/gkv007
 - [21] Mohan, Shikha, and Damodar Gupta. “Crosstalk of toll-like receptors signaling and Nrf2 pathway for regulation of inflammation.” *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* vol. 108 (2018): 1866-1878. doi:10.1016/j.biopha.2018.10.019
 - [22] Mahmoudvand, Hossein et al. “Toxoplasma gondii Infection Potentiates Cognitive Impairments of Alzheimer's Disease in the BALB/c Mice.” *The Journal of parasitology* vol. 102,6 (2016): 629-635. doi:10.1645/16-28
 - [23] White, Mitchell R et al. “Alzheimer's associated β -amyloid protein inhibits influenza A virus and modulates viral interactions with phagocytes.” *PloS one* vol. 9,7 e101364. 2 Jul. 2014, doi:10.1371/journal.pone.0101364
 - [24] Nasrabady, Sara E et al. “White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes.” *Acta neuropathologica communications* vol. 6,1 22. 2 Mar. 2018, doi:10.1186/s40478-018-0515-3
 - [25] Carbonell F, Charil A, Zijdenbos AP, Evans AC, Bedell BJ; Alzheimer's Disease Neuroimaging Initiative. Hierarchical multivariate covariance analysis of metabolic connectivity. *J Cereb Blood Flow Metab.* 2014 Dec;34(12):1936-43. doi: 10.1038/jcbfm.2014.165. Epub 2014 Oct 8. PMID: 25294129; PMCID: PMC4269748.
 - [26] D'Antonio F, Di Vita A, Zazzaro G, Brusà E, Trebbastoni A, Campanelli A, Ferracuti S, de Lena C, Guariglia C, Boccia M. Psychosis of Alzheimer's disease: Neuropsychological and neuroimaging longitudinal study. *Int J Geriatr Psychiatry.* 2019 Nov;34(11):1689-1697. doi: 10.1002/gps.5183. Epub 2019 Aug 14. PMID: 31368183.
 - [27] Liu, Z., “Dysfunctional whole brain networks in mild cognitive impairment patients: an fMRI study”, in *Medical Imaging 2012: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 2012, vol. 8317. doi:10.1117/12.910864.
 - [28] Narayan PJ, Lill C, Faull R, Curtis MA, Dragunow M. Increased acetyl and total histone levels in post-mortem Alzheimer's disease brain. *Neurobiol Dis.* 2015 Feb;74:281-94. doi: 10.1016/j.nbd.2014.11.023. Epub 2014 Dec 5. PMID: 25484284.
 - [29] Cooley SA, Cabeen RP, Laidlaw DH, Conturo TE, Lane EM, Heaps JM, Bolzenius JD, Baker LM, Salminen LE, Scott SE, Paul RH. Posterior brain white matter abnormalities in older adults with probable mild cognitive impairment. *J Clin Exp Neuropsychol.* 2015;37(1):61-9. doi: 10.1080/13803395.2014.985636. Epub 2014 Dec 18. PMID: 25523313; PMCID: PMC4355053.
 - [30] Jensen MM, Arvaniti M, Mikkelsen JD, Michalski D, Pinborg LH, Härtig W, Thomsen MS. Prostate stem cell antigen interacts with nicotinic acetylcholine receptors and is affected in Alzheimer's disease. *Neurobiol Aging.* 2015 Apr;36(4):1629-1638. doi: 10.1016/j.neurobiolaging.2015.01.001. Epub 2015 Jan 7. PMID: 25680266.
 - [31] Han P, Caselli RJ, Baxter L, Serrano G, Yin J, Beach TG, Reiman EM, Shi J. Association of pituitary adenylate cyclase-activating polypeptide with cognitive decline in mild cognitive impairment due to Alzheimer disease. *JAMA Neurol.* 2015 Mar;72(3):333-9. doi: 10.1001/jamaneurol.2014.3625. PMID: 25599520; PMCID: PMC5924703.
 - [32] Villeneuve S, Rabinovici GD, Cohn-Sheehy BI, Madison C, Ayakta N, Ghosh PM, La Joie R,

- Arthur-Bentil SK, Vogel JW, Marks SM, Lehmann M, Rosen HJ, Reed B, Olichney J, Boxer AL, Miller BL, Borys E, Jin LW, Huang EJ, Grinberg LT, DeCarli C, Seeley WW, Jagust W. Existing Pittsburgh Compound-B positron emission tomography thresholds are too high: statistical and pathological evaluation. *Brain*. 2015 Jul;138(Pt 7):2020-33. doi: 10.1093/brain/awv112. Epub 2015 May 6. PMID: 25953778; PMCID: PMC4806716.
- [33] Willette AA, Modanlo N, Kapogiannis D; Alzheimer's Disease Neuroimaging Initiative. Insulin resistance predicts medial temporal hypermetabolism in mild cognitive impairment conversion to Alzheimer disease. *Diabetes*. 2015 Jun;64(6):1933-40. doi: 10.2337/db14-1507. Epub 2015 Jan 9. PMID: 25576061; PMCID: PMC4439566.
- [34] Willette AA, Modanlo N, Kapogiannis D; Alzheimer's Disease Neuroimaging Initiative. Insulin resistance predicts medial temporal hypermetabolism in mild cognitive impairment conversion to Alzheimer disease. *Diabetes*. 2015 Jun;64(6):1933-40. doi: 10.2337/db14-1507. Epub 2015 Jan 9. PMID: 25576061; PMCID: PMC4439566.
- [35] Redolfi A, Manset D, Barkhof F, Wahlund LO, Glatard T, Mangin JF, Frisoni GB; neuGRID Consortium, for the Alzheimer's Disease Neuroimaging Initiative. Head-to-head comparison of two popular cortical thickness extraction algorithms: a cross-sectional and longitudinal study. *PLoS One*. 2015 Mar 17;10(3):e0117692. doi: 10.1371/journal.pone.0117692. PMID: 25781983; PMCID: PMC4364123.
- [36] Yamashita KI, Taniwaki Y, Utsunomiya H, Taniwaki T. Cerebral blood flow reduction associated with orientation for time in amnesic mild cognitive impairment and Alzheimer disease patients. *J Neuroimaging*. 2014 Nov-Dec;24(6):590-594. doi: 10.1111/jon.12096. Epub 2014 Mar 5. PMID: 24593247.
- [37] Ramos P, Santos A, Pinto NR, Mendes R, Magalhães T, Almeida A. Anatomical regional differences in selenium levels in the human brain. *Biol Trace Elem Res*. 2015 Feb;163(1-2):89-96. doi: 10.1007/s12011-014-0160-z. Epub 2014 Nov 21. PMID: 25413879.
- [38] Lu AT, Hannon E, Levine ME, Hao K, Crimmins EM, Lunnon K, Kozlenkov A, Mill J, Dracheva S, Horvath S. Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum. *Nat Commun*. 2016 Feb 2;7:10561. doi: 10.1038/ncomms10561. PMID: 26830004; PMCID: PMC4740877.
- [39] Sherva R, Tripodis Y, Bennett DA, Chibnik LB, Crane PK, de Jager PL, Farrer LA, Saykin AJ, Shulman JM, Naj A, Green RC; GENAROAD Consortium; Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Genetics Consortium. Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimers Dement*. 2014 Jan;10(1):45-52. doi: 10.1016/j.jalz.2013.01.008. Epub 2013 Mar 25. PMID: 23535033; PMCID: PMC3760995.
- [40] Khondoker M, Newhouse S, Westman E, Muehlboeck JS, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Lovestone S, Dobson R, Simmons A; AddNeuroMed consortium; Alzheimer's Disease Neuroimaging Initiative. Linking Genetics of Brain Changes to Alzheimer's Disease: Sparse Whole Genome Association Scan of Regional MRI Volumes in the ADNI and AddNeuroMed Cohorts. *J Alzheimers Dis*. 2015;45(3):851-64. doi: 10.3233/JAD-142214. PMID: 25649652.
- [41] Moon SW, Dinov ID, Kim J, Zamanyan A, Hobel S, Thompson PM, Toga AW. Structural Neuroimaging Genetics Interactions in Alzheimer's Disease. *J Alzheimers Dis*. 2015;48(4):1051-63. doi: 10.3233/JAD-150335. PMID: 26444770; PMCID: PMC4730943.

- [42] Leong W, Xu W, Wang B, Gao S, Zhai X, Wang C, Gilson E, Ye J, Lu Y. PP2A subunit PPP2R2C is downregulated in the brains of Alzheimer's transgenic mice. *Aging* (Albany NY). 2020 Apr 14;12(8):6880-6890. doi: 10.18632/aging.103048. Epub 2020 Apr 14. PMID: 32291379; PMCID: PMC7202491.
- [43] Silva B, Niehage C, Maglione M, Hoflack B, Sigrist SJ, Wassmer T, Pavlowsky A, Preat T. Interactions between amyloid precursor protein-like (APPL) and MAGUK scaffolding proteins contribute to appetitive long-term memory in *Drosophila melanogaster*. *J Neurogenet*. 2020 Mar;34(1):92-105. doi: 10.1080/01677063.2020.1712597. Epub 2020 Jan 22. PMID: 31965876.
- [44] Maphis NM, Jiang S, Binder J, Wright C, Gopalan B, Lamb BT, Bhaskar K. Whole Genome Expression Analysis in a Mouse Model of Tauopathy Identifies MECP2 as a Possible Regulator of Tau Pathology. *Front Mol Neurosci*. 2017 Mar 17;10:69. doi: 10.3389/fnmol.2017.00069. PMID: 28367114; PMCID: PMC5355442.
- [45] Zou Y, Zhang WF, Liu HY, Li X, Zhang X, Ma XF, Sun Y, Jiang SY, Ma QH, Xu DE. Structure and function of the contactin-associated protein family in myelinated axons and their relationship with nerve diseases. *Neural Regen Res*. 2017 Sep;12(9):1551-1558. doi: 10.4103/1673-5374.215268. PMID: 29090003; PMCID: PMC5649478.
- [46] Patak J, Hess JL, Zhang-James Y, Glatt SJ, Faraone SV. SLC9A9 Co-expression modules in autism-associated brain regions. *Autism Res*. 2017 Mar;10(3):414-429. doi: 10.1002/aur.1670. Epub 2016 Jul 21. PMID: 27439572.
- [47] Sathyan, Sanish et al. "Genetic Insights Into Frailty: Association of 9p21-23 Locus With Frailty." *Frontiers in medicine* vol. 5 105. 1 May. 2018, doi:10.3389/fmed.2018.00105
- [48] Mitjans, M et al. "Sexual dimorphism of AMBRA1-related autistic features in human and mouse." *Translational psychiatry* vol. 7,10 e1247. 10 Oct. 2017, doi:10.1038/tp.2017.213
- [49] Koga, Arthur T et al. "Genome-wide association analysis to predict optimal antipsychotic dosage in schizophrenia: a pilot study." *Journal of neural transmission* (Vienna, Austria : 1996) vol. 123,3 (2016): 329-38. doi:10.1007/s00702-015-1472-7
- [50] Nguyen, Thanh-Tung et al. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests." *BMC genomics* vol. 16 Suppl 2,Suppl 2 (2015): S5. doi:10.1186/1471-2164-16-S2-S5
- [51] Frost B. Alzheimer's disease: An acquired neurodegenerative laminopathy. *Nucleus*. 2016 May 3;7(3):275-83. doi: 10.1080/19491034.2016.1183859. Epub 2016 May 11. Erratum in: Extra view to: Frost B., , Bardai F.H., , Feany M.B. 2016. Lamin dysfunction mediates neurodegeneration in taupathies. *Current Biology* 26(1):129–136. <http://dx.doi.org/10.1016/j.cub.2015.11.039>. PMID: 27167528; PMCID: PMC4991240.
- [52] Wheeler JM, McMillan P, Strovast TJ, Liachko NF, Amlie-Wolf A, Kow RL, Klein RL, Szot P, Robinson L, Guthrie C, Saxton A, Kanaan NM, Raskind M, Peskind E, Trojanowski JQ, Lee VMY, Wang LS, Keene CD, Bird T, Schellenberg GD, Kraemer B. Activity of the poly(A) binding protein MSUT2 determines susceptibility to pathological tau in the mammalian brain. *Sci Transl Med*. 2019 Dec 18;11(523):eaao6545. doi: 10.1126/scitranslmed.aao6545. PMID: 31852801; PMCID: PMC7311111.
- [53] Huang NQ, Jin H, Zhou SY, Shi JS, Jin F. TLR4 is a link between diabetes and Alzheimer's disease. *Behav Brain Res*. 2017 Jan 1;316:234-244. doi: 10.1016/j.bbr.2016.08.047. Epub 2016 Aug 31. PMID: 27591966.

- [54] Kou X, Chen D, Chen N. Physical Activity Alleviates Cognitive Dysfunction of Alzheimer's Disease through Regulating the mTOR Signaling Pathway. *Int J Mol Sci.* 2019 Mar 29;20(7):1591. doi: 10.3390/ijms20071591. PMID: 30934958; PMCID: PMC6479697.
- [55] Tao Y, Han Y, Yu L, Wang Q, Leng SX, Zhang H. The Predicted Key Molecules, Functions, and Pathways That Bridge Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). *Front Neurol.* 2020 Apr 3;11:233. doi: 10.3389/fneur.2020.00233. PMID: 32308643; PMCID: PMC7145962.
- [56] Cruz-Rivera, Yazeli E et al. "A Selection of Important Genes and Their Correlated Behavior in Alzheimer's Disease." *Journal of Alzheimer's disease : JAD* vol. 65,1 (2018): 193-205. doi:10.3233/JAD-170799
- [57] Donlon, Timothy A, and Brian J Morris. "In silico analysis of human renin gene-gene interactions and neighborhood topologically associated domains suggests breakdown of insulators contribute to ageing-associated diseases." *Biogerontology* vol. 20,6 (2019): 857-869. doi:10.1007/s10522-019-09834-1
- [58] Abel, Mackenzie E et al. "KEOPS complex expression in the frontal cortex in major depression and schizophrenia." *The world journal of biological psychiatry : the official journal of the World Federation of Societies of Biological Psychiatry*, 1-10. 29 Sep. 2020, doi:10.1080/15622975.2020.1821917