frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fgene.2021.706986

Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression

Fengchun Ke†, Wei Kong† and Shuaiqun Wang*

College of Information Engineering, Shanghai Maritime University, Shanghai, China

Imaging genetics combines neuroimaging and genetics to assess the relationships between genetic variants and changes in brain structure and metabolism. Sparse canonical correlation analysis (SCCA) models are well-known tools for identifying meaningful biomarkers in imaging genetics. However, most SCCA models incorporate only diagnostic status information, which poses challenges for finding disease-specific biomarkers. In this study, we proposed a multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model to reveal disease-specific associations between single nucleotide polymorphisms and quantitative traits derived from multi-modal neuroimaging data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. MT-SCCAR uses complementary information carried by multiple-perspective cognitive scores and encourages group sparsity on genetic variants. In contrast with two other multi-modal SCCA models, MT-SCCAR embedded more accurate neuropsychological assessment information through linear regression and enhanced the correlation coefficients, leading to increased identification of high-risk brain regions. Furthermore, MT-SCCAR identified primary genetic risk factors for Alzheimer's disease (AD), including rs429358, and found some association patterns between genetic variants and brain regions. Thus, MT-SCCAR contributes to deciphering genetic risk factors of brain structural and metabolic changes by identifying potential risk biomarkers.

Keywords: imaging genetics, sparse canonical correlation analysis, magnetic resonance imaging, positron emission tomography, single nucleotide polymorphisms, multi-task learning

OPEN ACCESS

Edited by:
Jie Li,
Harbin Institute of Technology, China

Reviewed by:
Lei Du,
Northwestern Polytechnical University, China
Jin Li,
Harbin Medical University, China

***Correspondence:**
Wei Kong
weikong@shmtu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *08 May 2021*
Accepted: *19 July 2021*
Published: *05 August 2021*

Citation:
Ke F, Kong W and Wang S (2021) Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression. Front. Genet. 12:706986. doi: 10.3389/fgene.2021.706986

INTRODUCTION

Imaging genetics has recently emerged as a method for investigating imaging and genetic biomarkers related to diseases such as Alzheimer's disease (AD) (Bogdan et al., 2017). Identified neuroimaging and genetics biomarkers can provide a complementary understanding of the brain's structure and metabolism (Zhang et al., 2011). Moreover, the vast amounts of diagnostic and neuropsychological information from various perspectives enable the discovery of disease-specific biomarkers. Therefore, it is essential to simultaneously analyze multiple neuroimaging techniques, such as magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET), genotyping, and clinical diagnostic data. In this study, we aimed to build a model to identify disease-specific biomarkers across multiple imaging modalities, which can be used as an effective clue for disease diagnosis and targeted therapy.

Numerous studies have attempted to identify the associations between genotypic data such as single nucleotide polymorphisms (SNPs) and neuroimaging quantitative traits (QTs) (Rasetti and Weinberger, 2011). Because genotypic data and imaging QTs are multivariate, several bi-multivariate methods have been proposed to better characterize their associations. Liu et al. explored parallel independent component analysis (PICA) to detect the associations between brain function and genetic variants. However, this method cannot restore meaningful SNPs and regions of interest (ROIs), which has led to a lack of reasonable biomarker interpretation (Liu et al., 2009). Sparse canonical correlation analysis (SCCA) has a strong capability for bi-multivariate association identification and interpretable variable selection. Accordingly, many efforts have attempted to apply SCCA to neuroimaging genetics. Boutte et al. introduced an SCCA model with least absolute shrinkage and selection operator (LASSO) constraints on neuroimaging genetics data fusion (Boutte and Liu, 2010). Hao et al. presented a multi-view SCCA model to establish associations between SNPs, QTs, and cognitive outcomes (Hao et al., 2017). However, these multi-view SCCA models are a simple extension to conventional SCCA models. The requirement that SNP canonical weight vectors associate with all modal data is too strict, and could result in not making full use of all modal information. To address this limitation, Du et al. developed a multi-task SCCA model that could be used to jointly analyze SNPs and multiple neuroimaging data by treating each association as an individual learning task (Du et al., 2021). However, this model's neglect of diagnostic information means that biomarkers identified by these multiple-data models may not be sufficiently disease-specific.

To detect more complex and meaningful associations, studies to date have applied diagnostic information into SCCA methods (Yan et al., 2018; Du et al., 2020). Yan et al. proposed an outcome-relevant SCCA model based on a subject similarity matrix (Yan et al., 2018). Du et al. integrated multi-task SCCA and logistic regression in a sophisticated model to identify robust disease-related imaging and genetic patterns by incorporating diagnostic status information (Du et al., 2020). Classified diagnostic information, such as AD, mild cognitive impairment (MCI), and healthy control (HC), facilitates the association between SNPs and QTs; however, roughly dividing the disease stages does not provide any more accurate information than do continuous neuropsychological assessments measured from different angles.

To address the above problems, we proposed a novel SCCA model with the capacity to extract disease-specific biomarkers across multiple neuroimaging modalities. The proposed multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model integrates multi-task SCCA and multi-task linear regression in a fused model and uses multiple cognitive scores (CSs) as auxiliary information to induce associations between SNPs and QTs. Multi-task sparse canonical correlation analysis and regression considers the relationships within subjects from different disease courses and can find disease-specific biomarkers. We also considered underlying hierarchical information among SNPs by modeling structural relationships as divided by gene or by linkage disequilibrium (LD) in a group sparsity penalty. To evaluate MT-SCCAR's effectiveness, we performed extensive experiments to find associations between SNPs and two imaging QTs, including gray matter density and standard uptake value ratio (SUVR) extracted from MRI and positron emission tomography (PET), respectively. Compared with the other two multi-modal SCCA models that used real Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort data, MT-SCCAR not only outperformed these models in its ability to identify genetic AD risk factors, but also detected robust AD brain risk regions across multiple neuroimaging modalities. Thus, our proposed model has the potential to understand disease mechanisms from both structural and metabolic perspectives.

MATERIALS AND METHODS

Data Sources and Preprocessing

Real neuroimaging and genetic data used in this study were obtained from the ADNI1 database. A total of 305 non-Hispanic Caucasian subjects with genotype, neuroimaging, and cognitive assessment data at the ADNI1 baseline were downloaded from the LONI website,[1] including 83 HC, 148 MCI, and 74 AD subjects. The Mini-Mental State Examination (MMSE) is a numeric scale to test cognitive functions, including attention, calculation, and responsiveness to simple commands (Tombaugh and McIntyre, 1992). The Functional Activities Questionnaire (FAQ) evaluates instrumental activities of daily life, such as financial management and meal preparation (Teng et al., 2010). The Alzheimer's Disease Assessment Scale Cognitive Subscale (ADAS-Cog) mainly measures cognitive ability such as word recall, comprehension of spoken language, and orientation (Cano et al., 2010). **Table 1** shows the characteristics of the subjects.

Genotyping Data and Processing

Genotypes for 305 subjects were performed using the Illumina HumanHap610-Quad BeadChips from the ADNI1 database. The SNP data were lifted to hg19 build using lift over tool (Kent et al., 2002). To get pure SNP data, we used a genetic analysis tool PLINK (Purcell et al., 2007) to filter the SNPs using the following quality control criteria: gender check, sibling pair identification, call rate check (<90%) per subject and SNP marker, the Hardy-Weinberg Equilibrium (HWE $p < 10–6$), and marker removal by the minor allele frequency (MAF < 0.05), SNP data were further imputed using Michigan imputation server to estimate

[1] http://adni.loni.usc.edu/

TABLE 1 | Characteristics of the subjects.

Subjects	HC	MCI	AD
Number	83	148	74
Gender(M/F)	50/33	98/50	39/35
Age(mean ± std)	77.76 ± 4.59	76.62 ± 6.92	76.96 ± 6.91
Education(mean ± std)	15.68 ± 3.09	15.88 ± 2.77	14.27 ± 3.37
MMSE (mean ± std)	29.18 ± 1.11	26.09 ± 3.14	20.57 ± 3.20
FAQ (mean ± std)	0.54 ± 1.25	6.62 ± 8.96	19.75 ± 3.50
ADAS-Cog(mean ± std)	6.00 ± 2.89	13.72 ± 3.03	26.09 ± 11.64

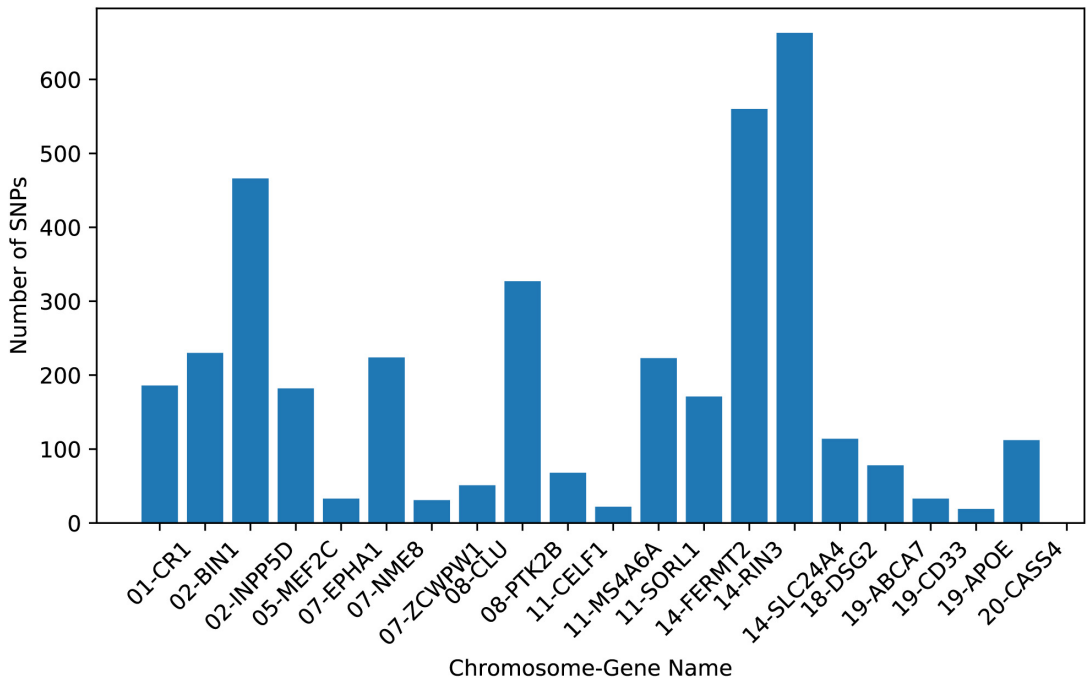

FIGURE 1 | The numbers of SNPs belonging to each AD risk gene used in this study.

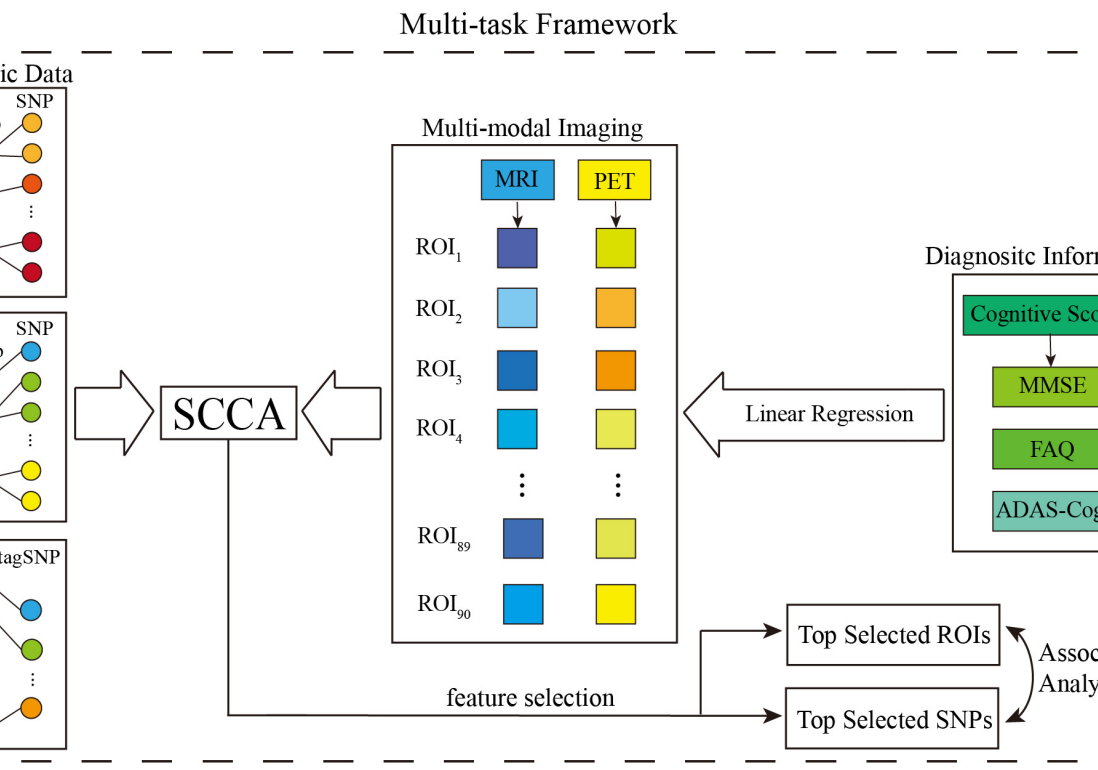

FIGURE 2 | Schematic illustration of MT-SCCAR.

the missing genotypes based on the HRC r1.1 2016 panel (Das et al., 2016). The post-imputation quality control used the rsq > 0.3 and MAF of 0.1 (Li et al., 2010).

Since our study focused on the top 20 AD risk genes listed on the AlzGene database[2] and references (Tanzi et al., 2007; Wang et al., 2012a). After imputation, we selected all the SNPs within ± 5k base pairs of the gene boundary using the ANNOVAR annotation (Wang et al., 2010). The above procedures yielded 3793 SNPs belonging to the top 20 risk genes. **Figure 1** presents the AD risk genes and the number of pre-selected SNPs. Moreover, considering the structural relationship among SNPs, we used Haploview (Barrett et al., 2004) to divide the LD block using the LD-Spline algorithm with $D^{'} > 0.8$, resulting in 209 blocks containing 3770 SNPs. A total of 894 tag SNPs were also assigned by Haploview in pairwise mode and an r^2 threshold was set to 0.8. These tagged SNPs represented the genetic variation across a particular region and could facilitate the association study (Montpetit et al., 2006). Furthermore, each SNP value was coded in an additive fashion to reflect the number of minor alleles.

Neuroimaging Data and Processing

The baseline 1.5T MRI scans were aligned to the standard Montreal Neurological Institute (MNI) space, resampled to $2 \times 2 \times 2$ mm^3 voxels, registered by SPM software package (Ashburner and Friston, 2007). Then, we extracted the gray matter tissue from the MRI scans and calculated mean gray matter densities of 116 ROIs based on MarsBar AAL atlas (Tzourio-Mazoyer et al., 2002). After removing 26 ROIs of the cerebellum, mean gray matter densities of 90 ROIs were used as QTs in our study.

The FDG-PET scans were co-registered to each subject's same visit MRI scans and normalized to MNI space by SPM tool. We further excluded white matter regions by masking the PET with gray matter masks obtained by the segmentation of the same subject's co-registered MRI. Then, the PET scans were normalized into the cerebellar gray matter reference region defined on the AAL atlas to generate SUVR images. After this, we used SUVR of 90 ROIs as QTs in our study by removing the 26 ROIs of cerebellum. Moreover, all the QTs were adjusted to exclude the influence of gender, age, and education.

Methods

In this paper, we denote lowercase letters as vectors, uppercase letters as matrices. $||\boldsymbol{x}||_2$ denotes the Euclidean norm, $||\boldsymbol{X}||_{2,1}$ denotes the sum of the Euclidean norms of the rows of $\boldsymbol{X}$, and $||\boldsymbol{X}||_{1,1}$ denotes the absolute sum of all elements of X.

The CS-Related Features Selection Model for Imaging Genetics

Assuming that there are n subjects with pSNPs, qROIs from M imaging modalities, and G different cognitive outcomes. We used $\boldsymbol{X} \in R^{n \times p}$,$\boldsymbol{Y}_m \in R^{n \times q}\,(m = 1, ..., M)$, and $\boldsymbol{z}_g \in R^{n \times 1}\,(g = 1, ..., G)$ to represent genetic data, multiple imaging data, and cognitive scores, respectively. The basic principle of MT-SCCAR is to find $\boldsymbol{U} \in R^{p \times M}$ and $\boldsymbol{V} \in R^{q \times M}$ to maximize the correlation between $\boldsymbol{X}\boldsymbol{u}_m$ and $\boldsymbol{Y}_m\boldsymbol{u}_m$, where u_{im} indicates the weight of the ith SNP for the mth modality, and v_{jm} indicates the weight of the jth ROI for the mth modality. To identify imaging genetic biomarkers that are relevant to CS and disease, the multi-task linear regression objective was combined with the multi-task SCCA (MTSCCA) objective, which can be formulated as:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{L}_R(\boldsymbol{V}) + \mathcal{L}_{SCCA}(\boldsymbol{U}, \boldsymbol{V}) + \Omega(\boldsymbol{U}) + \Omega(\boldsymbol{V}). \quad (1)$$

The above model consists of four parts, $\mathcal{L}_R(\boldsymbol{V})$ detects disease-relevant imaging QTs. $\mathcal{L}_{SCCA}(\boldsymbol{U}, \boldsymbol{V})$ captures the bi-multivariate associations between SNPs and multiple imaging QTs. $\Omega(\boldsymbol{U})$ and $\Omega(\boldsymbol{V})$ are the regularization terms to enforce sparsity of $\boldsymbol{U}$ and $\boldsymbol{V}$, so only a small number of interpretable variables can be selected. This model integrates the advantages of MTSCCA and linear regression, which has a certain superiority in using complementary cognitive information. **Figure 2** provides a schematic overview of MT-SCCAR. SNPs were classified into the same group by either gene or LD. Accordingly, SNPs with gene or LD information and tagSNPs were input to the SCCA component separately, which was used to establish the relationships between genetic data and multiple imaging data. The linear regression component was used to introduce CSs into the SCCA part. The multi-task modeling method guaranteed the ability to process multiple imaging and CS data. Unlike conventional unsupervised SCCA models, MT-SCCAR is a supervised SCCA model, which considers the relationships within subjects from different disease courses.

The Linear Regression Model for CS-QT Associations

In the proposed model, the associations between CSs and multi-modal neuroimaging QTs were established by multi-task regression. For each task, we built a regression model for revealing CS-related neuroimaging QTs:

$$\mathcal{L}_R(\boldsymbol{V}) = \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2, \quad (2)$$

TABLE 2 | Specific procedure of MT-SCCAR algorithm.

Algorithm: MT-SCCAR algorithm

Input: The genetic data $\boldsymbol{X} \in R^{n \times p}$, the neuroimaging data $\boldsymbol{Y} \in R^{n \times q}$ of M modalities, and the CS data $\boldsymbol{Z} \in R^{n \times C}$.
λ_{u1}, λ_{u2}, λ_{u3}, λ_{v1}, λ_{v2}, γ_u, and γ_v.
Ensure: canonical weights $\boldsymbol{V}$ and $\boldsymbol{U}$
1: While not converged regarding to $\boldsymbol{V}$, $\boldsymbol{U}$ do
2:Update the diagonal matrix $\boldsymbol{D}_{v1}$ and $\boldsymbol{D}_{v2}$;
3:Solve $\boldsymbol{v}_m$ according to Equation (12);
4:Normalize $\boldsymbol{v}_m$ so that $||\boldsymbol{Y}\boldsymbol{v}_m||_2^2 = 1$;
5:Update the diagonal matrix $\boldsymbol{D}_{u1}$, $\boldsymbol{D}_{u2}$ and $\boldsymbol{D}_{u3}$;
6:Solve $\boldsymbol{U}$ according to Equation (15);
7:Normalize $\boldsymbol{u}_m$ so that $||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1$;
8: End while

[2]www.alzgene.org

where M is the number of neuroimaging modalities, C is the number of cognitive assessments, and n is the total amount of subjects. $\boldsymbol{v}_m$ is the canonical weight of QTs for the mth modalities,$\boldsymbol{y}_m^l$ is the neuroimaging data vector of the lth subjects for the mth modalities, and z_c^l is the score of the lth subjects for the cth cognitive assessments. This multi-task regression model can jointly utilize neuropsychological assessments from different complementary perspectives.

The MTSCCA Model for SNP-QT Associations

Unlike conventional multi-view SCCA models, MTSCCA learns multiple SCCA tasks together by treating each imaging modality association model as a task. This model was proposed by Du et al. (2021) and can be defined as:

$$\min_{\boldsymbol{u}_m,\boldsymbol{v}_m} \sum_{m=1}^{M} -\boldsymbol{u}_m^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m \; s.t. \; ||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1, \; ||\boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 = 1, \; \forall m. \tag{3}$$

For canonical weights $\boldsymbol{U}$ and $\boldsymbol{V}$, each column $\boldsymbol{u}_m$ and $\boldsymbol{v}_m$ represents an individual learning task for different modalities. The main advantage of this multi-task strategy is that SNP canonical weight vectors do not need to be associated with all imaging modalities simultaneously. Each task focuses on identifying SNPs that are associated with only one imaging modality.

The Regularization Terms

Multiple neuroimaging modalities can provide more comprehensive information in terms of both structural and functional perspectives. In our model, two principal tasks corresponded to two neuroimaging modalities. MT-SCCAR should be able to identify neuroimaging QTs shared among multiple modalities and to enforce individual level sparsity. Hence, $\boldsymbol{\Omega}(\boldsymbol{V})$ was composed of two parts, which can be defined as:

$$\boldsymbol{\Omega}(\boldsymbol{V}) = \lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1}, \tag{4}$$

where λ_{v1} and λ_{v2} are positive parameters and can be tuned via cross-validation.

The first penalty was defined as:

$$||\boldsymbol{V}||_{2,1} = \sum_{i=1}^{q}\sqrt{\sum_{m=1}^{M}\boldsymbol{V}_{i,j}^2} = \sum_{i=1}^{q}\left|\left|\boldsymbol{V}_{i,:}\right|\right|_2, \tag{5}$$

This term aims to enforce task-consistent (modality-consistent) sparsity on $\boldsymbol{V}$, which encourages multi-modal imaging QTs to share similar canonical weights.

The second penalty was defined as:

$$||\boldsymbol{V}||_{1,1} = \sum_{j=1}^{q}\sum_{m=1}^{M}\left|v_{jm}\right|, \tag{6}$$

This term indicates the absolute sum of all elements of $\boldsymbol{V}$, which helps to screen the entire ROIs to find the relevant ROIs.

Similarly, the regularization terms of $\boldsymbol{U}$ also include the above two penalties, which can help discover SNPs that may affect multiple brain regions. It is common knowledge that some SNPs located in the same gene or LD block often have similar functions and are jointly related to specific ROIs. It is essential to model underlying hierarchical information among SNPs by adding an extra penalty. Therefore, we defined $\boldsymbol{\Omega}(\boldsymbol{U})$ as follows:

$$\boldsymbol{\Omega}(\boldsymbol{U}) = \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \lambda_{u3}||\boldsymbol{U}||_G, \tag{7}$$

where λ_{u1}, λ_{u2}, and λ_{u3} are positive parameters, the third penalty (Wang et al., 2012a) can be formulated as:

$$||\boldsymbol{U}||_G = \sum_{k=1}^{K}\sqrt{\sum_{i\in g_k}\sum_{j=1}^{M}u_{ij}^2}, \tag{8}$$

where K denotes the number of groups divided by gene or LD. This penalty penalizes canonical weights as a whole for each task and thus can fully use the structural information.

The Optimization Algorithm

In order to address the problem defined in Equation (1), according to the method that has been well studied previously (Du et al., 2021), we can rewrite Equation (1):

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V}} &\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2 + \\ &\sum_{m=1}^{M}||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \\ &\lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1} + \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \\ &\lambda_{u3}||\boldsymbol{U}||_G \; s.t. \; ||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1, \; ||\boldsymbol{Y}\boldsymbol{v}_m||_2^2 = 1, \; \forall m.\end{aligned} \tag{9}$$

We then use the Lagrange multiplier to solve this problem by taking the partial derivatives of Equation (9) regarding $\boldsymbol{u}_m$ and$\boldsymbol{v}_m$ separately, which can change the formula from non-convex to convex.

First, we treat $\boldsymbol{U}$ as constant, the Lagrange multiplier of Equation (9) can be simplified as:

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2 + \\ &\sum_{m=1}^{M}||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \\ &\lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1} + \gamma_v\sum_{m=1}^{M}||\boldsymbol{Y}_m\boldsymbol{v}_m||_2^2\end{aligned} \tag{10}$$

by dropping the constant terms, and γ_v is a positive parameter. For each $\boldsymbol{v}_m$, We further take the partial derivatives of Equation (10) and let the result be zero:

$$\begin{aligned}&\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m - \sum_{c=1}^{C}\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{z}_c - \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}_m + \lambda_{v1}\boldsymbol{D}_{v1}\boldsymbol{v}_m + \lambda_{v2}\boldsymbol{D}_{v2}\boldsymbol{v}_m \\ &+ (\gamma_v + 1)\,\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m = \boldsymbol{0},\end{aligned} \tag{11}$$

where $\boldsymbol{D}_{v1}$ is a diagonal matrix with the ith element as $\frac{1}{2||\boldsymbol{v}_{i,:}||_2}(i \in [1, q])$, and $\boldsymbol{D}_{v2}$ is a diagonal matrix with ith element as $\frac{1}{2||\boldsymbol{v}_{im}||_2}(i \in [1, q]$, and $m \in [1, M])$. Obviously, we can take an

iterative rule to solve this problem since both $\boldsymbol{D}_{v1}$ and $\boldsymbol{D}_{v2}$ are rely on canonical weights $\boldsymbol{V}$. This rule can be formulated as:

$$\boldsymbol{v}_m = \left(\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m + \lambda_{v1}\boldsymbol{D}_{v1} + \lambda_{v2}\boldsymbol{D}_{v2} + (\gamma_v + 1)\,\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\right)^{-1} \left(\sum_{c=1}^{C} \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{z}_c + \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}_m\right). \quad (12)$$

Then, we treat $\boldsymbol{V}$ as a constant, the Lagrange multiplier of Equation (9) can be simplified as:

$$\sum_{m=1}^{M} ||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \lambda_{u3}||\boldsymbol{U}||_G + \gamma_u\,||\boldsymbol{X}\boldsymbol{u}||_2^2 \quad (13)$$

by dropping the constant terms, and γ_u is also a positive parameter. Similar to $\boldsymbol{v}_m$, for $\boldsymbol{U}$, we let the partial derivatives of Equation (13) to be zero:

$$-\boldsymbol{X}^{\mathrm{T}}\mathbb{Y} + \lambda_{u1}\boldsymbol{D}_{u1}\boldsymbol{U} + \lambda_{u2}\boldsymbol{D}_{u2}\boldsymbol{U} + \lambda_{u3}\boldsymbol{D}_{u3}\boldsymbol{U} + \gamma_u\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}U = \boldsymbol{0}, \quad (14)$$

where $\boldsymbol{D}_{u1}$ is a diagonal matrix with the ith element as $\frac{1}{2||\boldsymbol{u}_{i,:}||_2}(i \in [1, p])$, $\boldsymbol{D}_{u2}$ is a diagonal matrix with ith element as $\frac{1}{2||\boldsymbol{u}_{im}||_2}(i \in [1, p]$, and $m \in [1, M])$, $\boldsymbol{D}_{u3}$ is a block diagonal matrix with element as $\frac{1}{2||U_{k,:}||_{\mathrm{F}}}\boldsymbol{I}_k(k \in [1, K])$, I_k is an identity matrix of the same size with kth SNP groups, and $\mathbb{Y} = [\boldsymbol{Y}_1\boldsymbol{v}_1\boldsymbol{Y}_2\boldsymbol{v}_2 \ldots \boldsymbol{Y}_m\boldsymbol{v}_m]$. Hence, the iterative rules can be formulated as:

$$\boldsymbol{U} = (\lambda_{u1}\boldsymbol{D}_{u1} + \lambda_{u2}\boldsymbol{D}_{u2} + \lambda_{u3}\boldsymbol{D}_{u3} + (\gamma_u + 1)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{Y}. \quad (15)$$

Based on the above analysis, the optimization algorithm of the proposed method is shown in **Table 2**. We can update $\boldsymbol{V}$ and $\boldsymbol{U}$ alternatively in each iteration until the predefined convergence criterion is satisfied.

RESULTS AND DISCUSSION

Experimental Settings

To comprehensively evaluate the effectiveness of our proposed MT-SCCAR model, two similar models that can analyze multi-modal data were compared with MT-SCCAR. They are three-view SCCA (TSCCA) and MTSCCA. Three-view SCCA can process neuroimaging, genetics, and cognitive scores data by extending conventional two-view association to three data types. MTSCCA was used to evaluate the regression part of our proposed model performance.

There are seven parameters in our model. Tuning all these parameters will pay a high cost. In our experiment, we fixed γ_u and γ_v to 1 since they mainly control the amplitude of $\boldsymbol{V}$ and $\boldsymbol{U}$ (Chen and Liu, 2011). To tune these

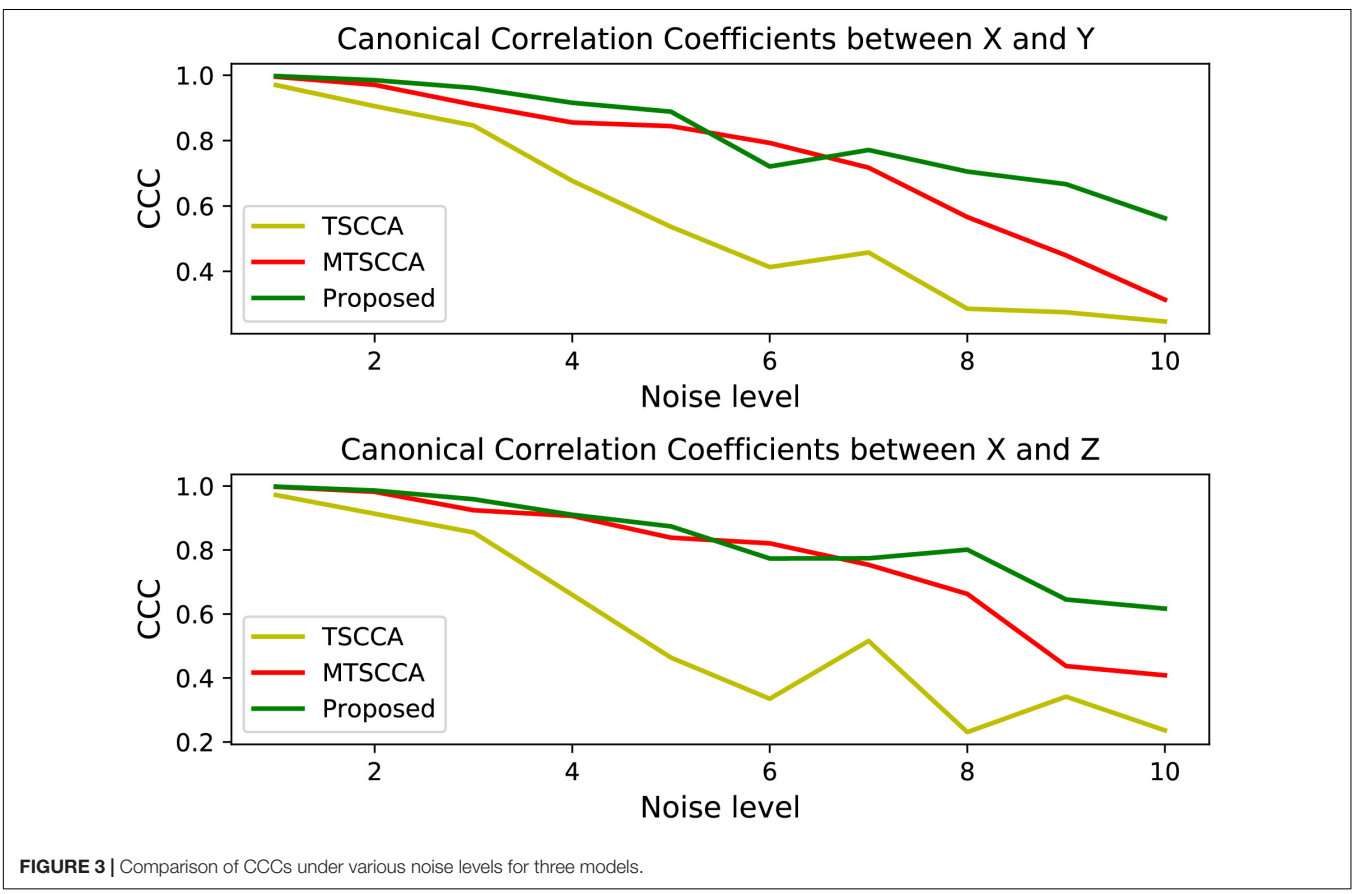

FIGURE 3 | Comparison of CCCs under various noise levels for three models.

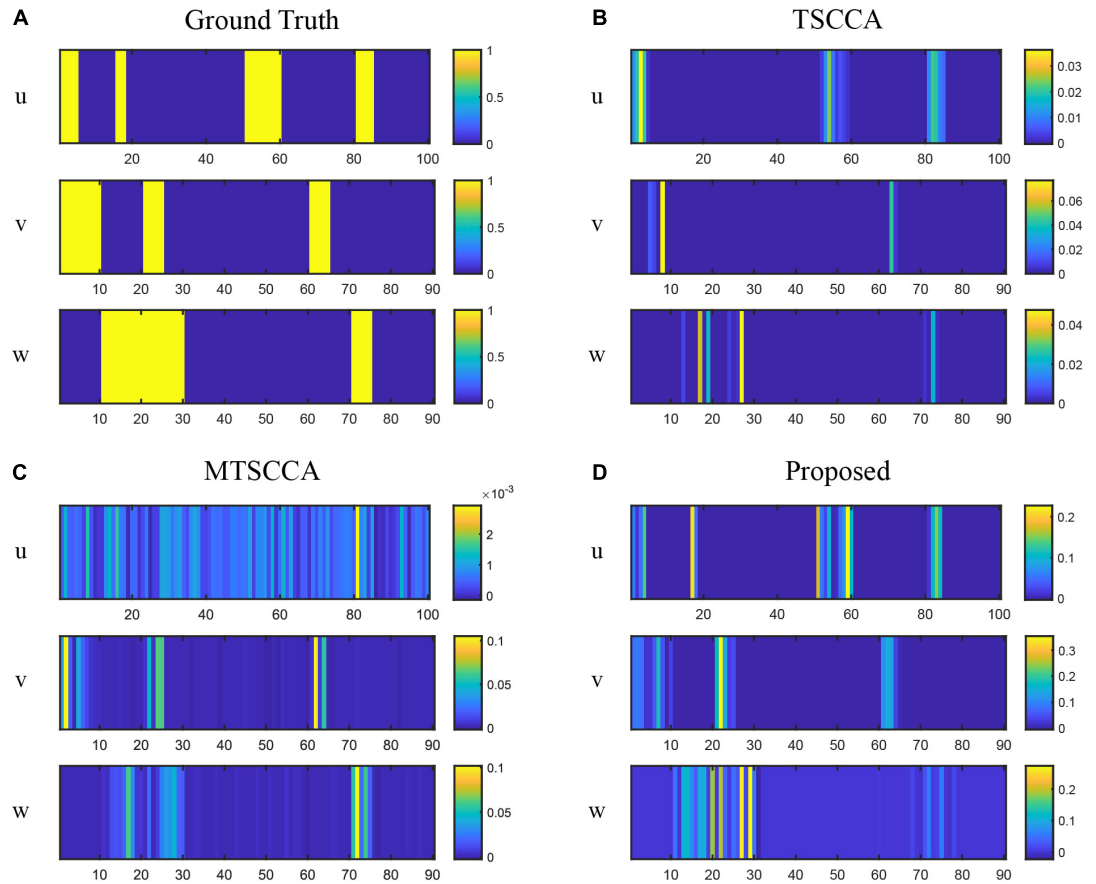

FIGURE 4 | Comparison of canonical weights on synthetic data with the high noise level. **(A)** The ground truth canonical weights. **(B)** The estimated canonical weights of TSCCA. **(C)** The estimated canonical weights of MTSCCA. **(D)** The estimated canonical weights of the proposed model.

parameters to appropriate values, we adopted a nested five-fold cross-validation strategy. Specifically, we tuned them in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$ until the highest mean testing canonical correlation coefficients (CCCs) was generated in the inner loop. CCC was defined as the Pearson correlation coefficient between $\boldsymbol{X}u$ and $\boldsymbol{Y}v$, and can be used as a quantitative measure of SCCA model performance (Hao et al., 2017). For multi-task learning, CCC can be calculated by $corr\ (\boldsymbol{X}_m\boldsymbol{u}_m, \boldsymbol{Y}_m\boldsymbol{v}_m)$ for mth task. Also, we terminated the iteration when both $\max\left|u_i^{(t+1)} - u_i^t\right| \leq 10^{-5}$ and $\max\left|v_j^{(t+1)} - v_j^t\right| \leq 10^{-5}$ were satisfied. All models in our experiment have taken the same parameter adjustment steps.

Results on Synthetic Data

We generated ten synthetic datasets with the same ground truth of loading vectors but different noise levels. Assuming that $\boldsymbol{X} \in R^{n \times p}$,$\boldsymbol{Y} \in R^{n \times q}$, and $\boldsymbol{Z} \in R^{n \times q}$ denote SNP, MRI, and PET for all synthetic data sets, respectively. $\boldsymbol{X}$ was generated by $\boldsymbol{X} = \boldsymbol{ul} + \boldsymbol{e}$,$\boldsymbol{Y}$ was generated by $\boldsymbol{Y} = \boldsymbol{vl} + \boldsymbol{e}$, and $\boldsymbol{Z}$ was generated by $\boldsymbol{Z} = \boldsymbol{wl} + \boldsymbol{e}$, where $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$are known loading vectors, $\boldsymbol{l}$ is a latent vector with a 3-component Gaussian distribution to simulate the disease course (Yan et al., 2018), and $\boldsymbol{e}$ is derived from the Gaussian distribution $N\left(0,\ \sigma_e^2\right)$ with σ_e^2 as the noise variance. In our study, n, p, and q were set to 90, 100, and 90, respectively. All the 90 samples were classed into three groups with centers -5, 0, 5. For neuropsychological assessment data, $\boldsymbol{c}$ was generated by $\boldsymbol{c} = \boldsymbol{l} + \boldsymbol{e}$. To assess the model performance at various noise levels, we tested different noise variances ranging from 1 to 10, with a step size of 1. The five-fold cross-validation results are shown in **Figures 3**, **4**.

Figure 3 plots the testing CCC for three models with changing noise levels. Higher CCC indicates better performance in identifying underlying associations. As expected, the performance decreased with increased noise levels for all models. All three models performed similarly well at low noise levels. Models with the multi-task framework (MTSCCA, MT-SCCAR) performed better than TSCCA at medium noise levels. Then MT-SCCAR outperformed the other two models as the noise level was further increased, suggesting that MT-SCCAR had a strong ability to resist noise. **Figure 4** shows the true signal of canonical weights and canonical weights estimated by three models with a noise level of 10. Important features were highlighted in the heatmaps displaying ground truth. We could clearly observe that the weight $\boldsymbol{u}$ estimated by MTSCCA was ambiguous. It was

therefore difficult to recognize important features. TSCCA did not identify complete important features. MT-SCCAR estimated the best canonical weights that were consistent with the ground truths. These results implied that the proposed model had the potential to extract important features in real neuroimaging genetics studies.

Results on Neuroimaging and Genetics Data

In real neuroimaging genetics data application, all subjects with SNP, MRI, PET, and three different cognitive information data were inputted into MT-SCCAR.A total of 3793 SNPs with LD or gene group information and 894 tag SNPs were used separately. The group sparsity penalty treated each tagSNP as an individual group. We then averaged the CCCs based on five-fold cross-validation, representing the mean strength of identified associations between SNPs and two imaging QTs.

As illustrated in **Table 3**, TSCCA achieved the highest training CCCs but performed poorly in testing CCCs. These unreasonable results may be caused by overfitting (Du et al., 2021). Multi-task sparse canonical correlation analysis and regression achieved the highest testing CCCs on both MRI and PET. Specifically, MT-SCCAR (LD) and MT-SCCAR (gene) achieved the highest testing CCC on SNP-MRI association and SNP-PET association, respectively. Notably, MT-SCCAR (gene) achieved relatively small testing CCC on SNP-MRI association; MT-SCCAR (LD) achieved a more balanced result than those of MT-SCCAR (gene), which indicates that using LD group information is more beneficial than using gene group information. The training CCCs of MT-SCCAR with tagSNP were higher than those of MT-SCCAR with group information since the different numbers

TABLE 3 | Comparison of canonical correlation coefficients (mean ± std) in terms of each model.

	Training CCCs		Testing CCCs	
	SNP-MRI	**SNP-PET**	**SNP-MRI**	**SNP-PET**
TSCCA	**0.82 ± 0.01**	**0.82 ± 0.01**	0.21 ± 0.05	0.23 ± 0.03
MTSCCA	0.55 ± 0.05	0.46 ± 0.11	0.21 ± 0.03	0.30 ± 0.06
Proposed (LD)	0.55 ± 0.01	0.48 ± 0.01	**0.34 ± 0.04**	0.36 ± 0.05
Proposed(gene)	0.56 ± 0.02	0.47 ± 0.01	0.22 ± 0.02	**0.39 ± 0.03**
Proposed(tagSNP)	0.60 ± 0.03	0.52 ± 0.01	0.26 ± 0.05	0.27 ± 0.04

The best correlation coefficients are shown in boldface.

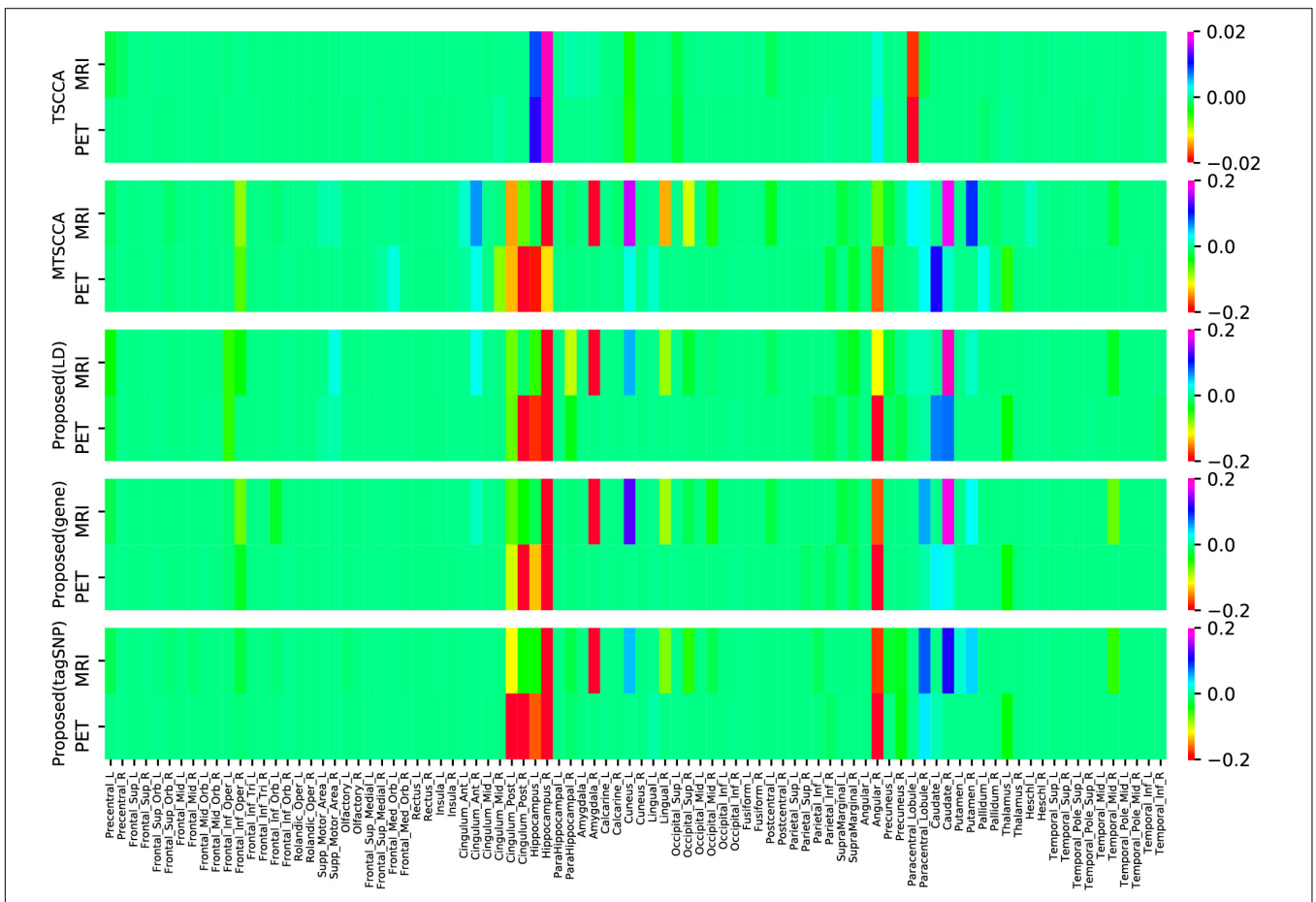

FIGURE 5 | Comparison of estimated canonical weights of imaging QTs. Each row represents: (1) TSCCA; (2) MTSCCA; (3) Proposed (LD); (4) Proposed (gene); (5) Proposed (tagSNP). Within each row, there are two parts represent two imaging modalities.

TABLE 4 | The top ten selected ROIs by the proposed model.

MRI	PET
Hippocampus_R	Cingulum_Post_R
Amygdala_R	**Angular_R**
Caudate_R	**Hippocampus_R**
Angular_R	**Hippocampus_L**
ParaHippocampal_R	**Caudate_R**
Lingual_R	Caudate_L
Cingulum_Post_L	**Cingulum_Post_L**
Cuneus_L	**Frontal_Inf_Oper_L**
Hippocampus_L	Thalamus_L
Frontal_Inf_Oper_L	**ParaHippocampal_R**

The jointly selected ROIs are shown in boldface.

of SNPs were used. Moreover, MTSCCA also performed better than TSCCA, which means the superiority of multi-task models when dealing with multiple imaging QTs and genetic data.

The Top Selected ROIs

In addition to the CCCs, the canonical weights were also one of the focuses of our study since they can help us find brain regions being highly related to AD. **Figure 5** shows the comparison of mean canonical weights of two imaging QTs based on five-fold cross-validation trials. Each row represents an SCCA model. The heatmap color represents the estimated weight of each model, so the selected QTs were highlighted in **Figure 5**. We can clearly observe that several brain regions were selected by both MRI and PET scans, such as the right hippocampal and the right angular gyrus, indicating that these regions may be modality-consistent. Additionally, TSCCA identified only modality-consistent QTs but failed to identify modality-specific QTs. This was due to the nature of its modeling strategy and may have resulted in crucial biomarkers being ignored. Multi-task models can identify modality-specific and modality-consistent QTs, which also implied the limitations of conventional multi-view SCCA models. In order to more accurately analyze the identified brain regions, using the proposed model with LD group information, the top ten ROIs of each modality were selected and sorted according to the absolute values of canonical weights.

As shown in **Table 4**, ROIs that were jointly selected by two imaging modalities are shown in boldface, all of which are known to be closely related to the pathogenesis of AD according to previous research. The hippocampus is essential for forming new memories and was reported as one of the earliest affected brain regions in AD and MCI (Moreno-Jimenez et al., 2019). Both left and right caudate nucleus have been reported that their volume is significantly different between AD and normal control (Cho et al., 2014; Botzung et al., 2019). The right angular gyrus is considered to be closely related to language ability, and patients with angular gyrus syndrome are often found to have damage in this brain

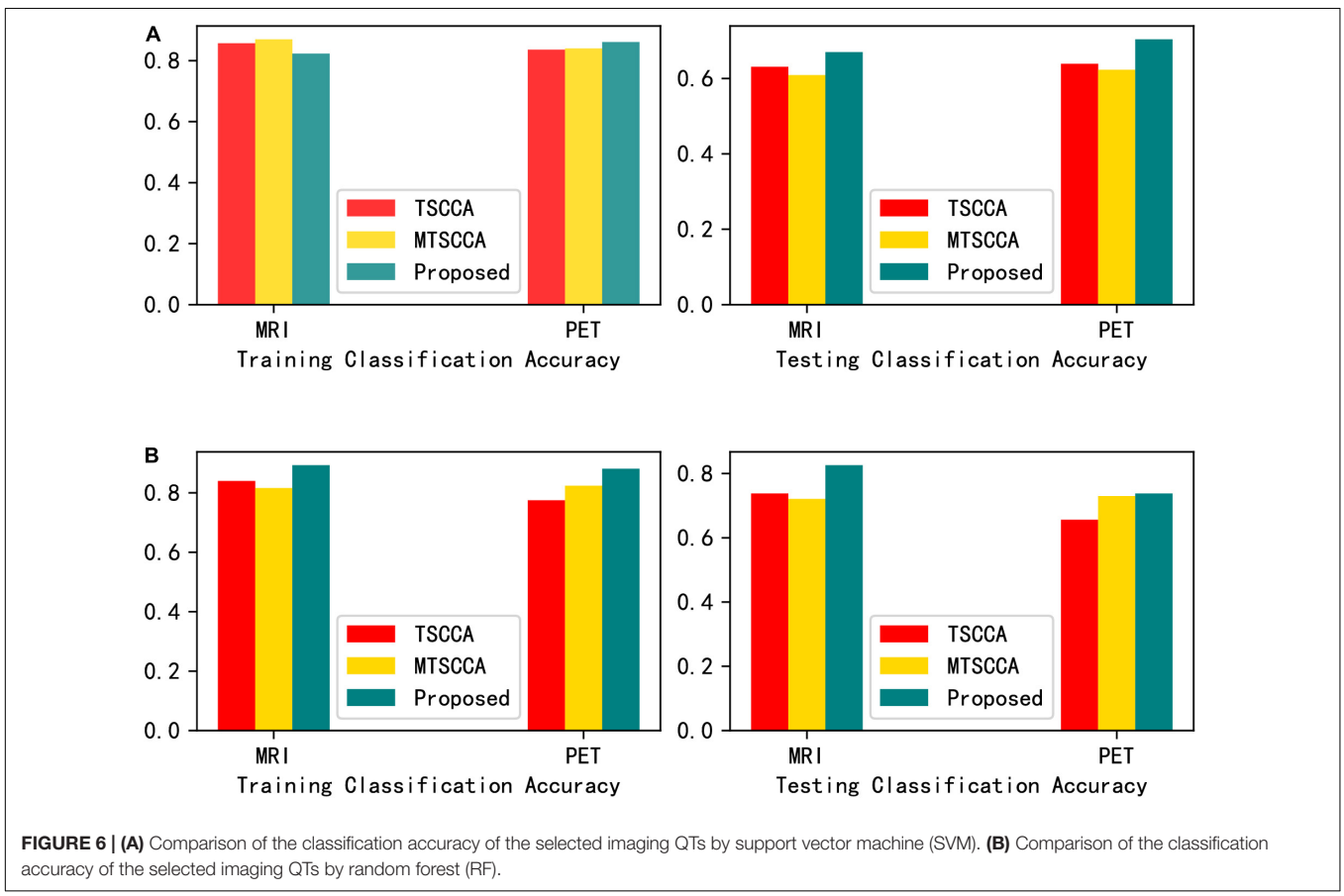

FIGURE 6 | (A) Comparison of the classification accuracy of the selected imaging QTs by support vector machine (SVM). **(B)** Comparison of the classification accuracy of the selected imaging QTs by random forest (RF).

TABLE 5 | The top ten selected SNPs.

TSCCA	MTSCCA	Proposed (LD)	Proposed (gene)	Proposed (tagSNP)
rs735780	rs769449	rs769449	rs7256200	rs117641527
rs405509	rs7256200	rs7256200	rs10414043	rs8012948
rs578506	rs10414043	rs10414043	rs769449	rs1884910
rs4904901	rs4904901	rs4901317	rs7157639	rs78015388
rs7157639	rs61975596	rs429358	rs405509	rs2598123
rs429358	rs7794735	rs4904901	rs4904901	rs4335936
rs4257390	rs55636820	rs7157639	rs429358	rs59325138
rs7412	rs77640937	rs449647	rs75773078	rs439401
rs7794735	rs34273097	rs11629428	rs11629428	rs112097633
rs10256195	rs9972149	rs3829947	rs4901317	rs429358

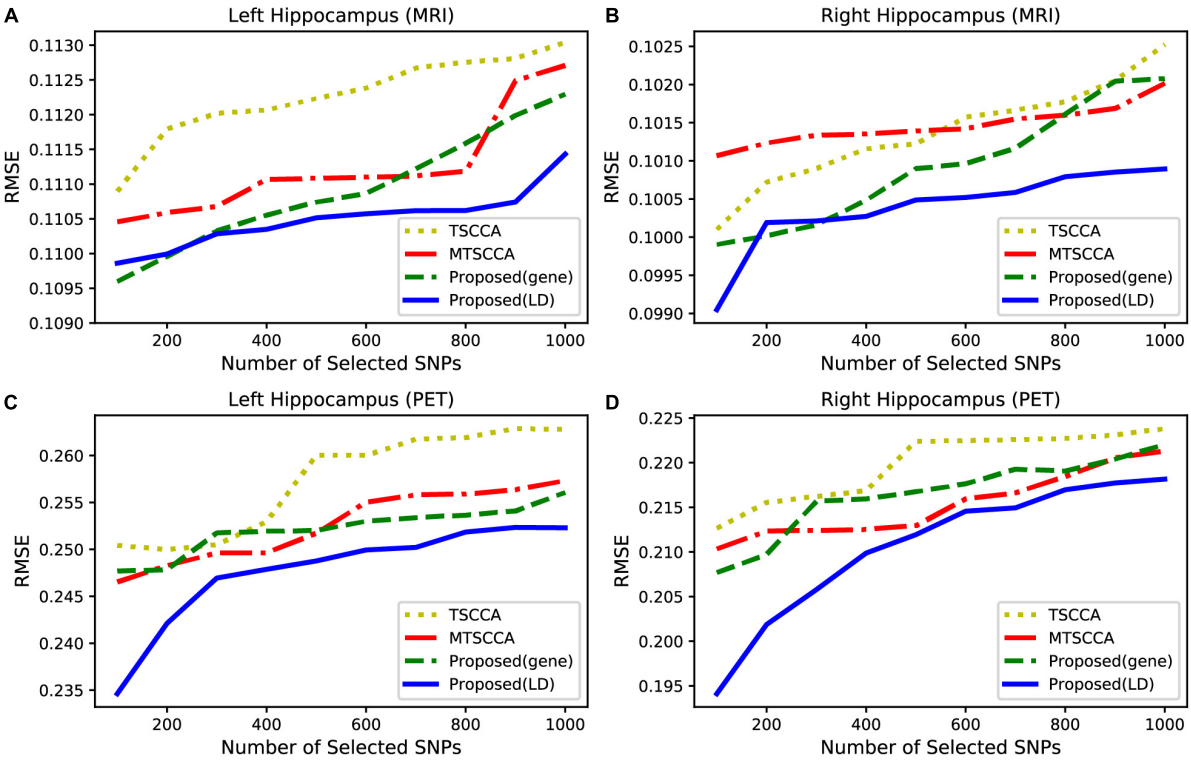

FIGURE 7 | Comparison of the RMSE with respect to different numbers of SNPs from 100 to 1000. **(A)** QT of left hippocampus based on MRI scan. **(B)** QT of right hippocampus based on MRI scan. **(C)** QT of left hippocampus based on PET scan. **(D)** QT of right hippocampus based on PET scan.

area (Horwitz et al., 1998). The right parahippocampal gyrus affects the encoding and maintenance of bound information related to working memory (Luck et al., 2010). The metabolic reduction in the posterior cingulate gyrus is a very early sign in AD (Minoshima et al., 1997). Notably, all the remaining brain regions have also been reported to be associated with AD in published literature. These satisfactory results were due to the inclusion of cognitive information into the linear regression to adjust weighting.

In order to further thoroughly verify that the neuroimaging biomarkers found by the proposed model are more disease-related than those found by the other two models. Selecting the top ten QTs as input features, support vector machine (SVM) with Gaussian radial basis function (RBF) kernel and random forest (RF) were adopted as classification methods. The parameters were tuned with five-fold cross-validation based on the training sets. **Figure 6** presents the classification accuracies of the two classifiers. The testing classification results showed that the classifier using the features selected by MT-SCCAR achieved the highest accuracies, thus indicating the superiority of MT-SCCAR in identifying disease-related biomarkers. Notably, the testing classification accuracies were relatively low for both SVM and RF, probably due to inevitable noise during the feature extraction process of brain imaging. These results were also consistent with previous studies (Wang et al., 2012b; Adeli et al., 2017).

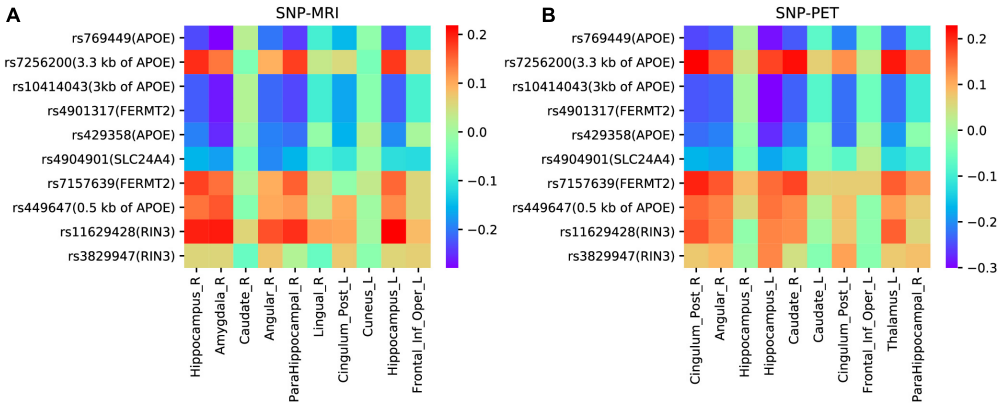

FIGURE 8 | (A) The heatmap of pairwise correlations between SNPs and ROIs of MRI scan. **(B)** The heatmap of pairwise correlations between SNPs and ROIs of PET scan.

TABLE 6 | The correlation coefficients and p-values of eight SNP-ROI pairs.

SNP-ROI pairs	Correlation coefficient	p-value
rs4904901-Angular_R(MRI)	−0.189	0.002
rs4904901- Angular_R(PET)	0.180	0.003
rs7157639-Hippocampus_R(MRI)	0.176	0.003
rs7157639-Cingulum_Post_R(PET)	0.204	0.001
rs11629428-Hippocampus_L(MRI)	0.218	0.0003
rs11629428- Cingulum_Post_R(PET)	0.171	0.004
rs3829947- Angular_R(MRI)	0.078	0.067
rs3829947- Hippocampus_L(PET)	0.135	0.025

The Top Selected SNPs

In addition to neuroimaging biomarkers, SCCA models can also identify genetic biomarkers. We averaged the SNP canonical weights into a single vector and selected the top ten SNPs. As illustrated in **Table 5**, the proposed model with LD or gene group information yielded meaningful results. For example, rs769449 (APOE) is located in promoter and enhancer areas for multiple brain tissues and is associated with AD (Liu et al., 2018). Moreover, the well-known AD risk biomarker rs429358 (APOE) was also identified by the proposed model, demonstrating its strong correlation ability. The remaining five SNPs of the proposed model, i.e., rs7256200 (3.3 kb of APOE), rs10414043 (3kb of APOE), rs4901317 (FERMT2), rs449647 (0.5 kb of APOE), and rs405509 (0.2 kb of APOE), have also been documented to increase the risk of AD in previous studies (Lin et al., 2017; Xiao et al., 2017). However, four selected SNPs have not yet been reported to be related to AD. They still need further research to confirm in the future. Next, we compared the top ten SNPs identified by MT-SCCAR (LD and gene) with the 894 tagSNPs. Interestingly, MT-SCCAR (LD) identified six tagSNPs (rs7256200, rs4901317, rs429358, rs7157639, rs449647, and rs3829947). Multi-task sparse canonical correlation analysis and regression (gene) identified five tagSNPs (rs7256200, rs7157639, rs405509, rs429358, and rs4901317). This implied that using tagSNP will reduce the number of SNPs that need to be analyzed and facilitate identifying significant SNPs. The proposed model with tagSNP also identified some significant SNPs. For example, rs59325138 (3.6 kb of APOE) has been reported to modify the cerebrospinal fluid apolipoprotein E protein levels (Cervantes et al., 2011). The Beta-Amyloid (1-42), an AD biomarker, is associated with rs439401 (1.8kb of APOE) (Xu et al., 2014). The TSCCA identified the rs4292358 and three other SNPs (rs405509, rs7412, and rs7794735) that have been reported previously (Arking et al., 2008; Ma et al., 2016; Zhen et al., 2017). The MTSCCA also identified four SNPs (rs769449, rs7256200, rs10414043, and rs7794735) but cannot identify rs429358. In summary, the proposed model was more accurate for identifying disease-specific genetic biomarkers than the other two models.

Alzheimer's disease (AD) usually first affects the hippocampus, resulting in cognitive decline and memory loss (Moreno-Jimenez et al., 2019). Therefore, when selecting the same number of features, the predictive effect of the QTs of the hippocampus can be used to evaluate model performance. Based on this analysis, we built a regression model to predict the QTs of the hippocampus from MRI and PET scans. Different numbers of SNPs were selected from 100 to 1000 with a step of 100. Using a support vector machine (SVR) with RBF kernel, we calculated the average root mean squared error (RMSE) for each model based on five-fold cross-validation. For a fair comparison, we only compared TSCCA, MTSCCA, MT-SCCAR (gene), and MT-SCCAR(LD) since MT-SCCAR (tagSNP) used only 894 tagSNPs. **Figure 7** shows the testing RMSE of the left and right hippocampus obtained by different imaging techniques. Smaller RMSE indicates that the selected SNPs are more related to AD. According to **Figure 7**, the prediction errors were lowest for the proposed model. These results suggested that the proposed model outperformed the other two models on four imaging QTs.

Pairwise Correlation Analyses

Based on the top ten selected ROIs and SNPs obtained by the proposed model with LD group information, we drew heatmaps of pairwise correlation coefficients between SNPs and two imaging QTs. As illustrated in **Figure 8**, it is clearly observed that

the selected SNPs were mainly located in and around the APOE region. APOE is the major genetic risk factor for AD (Munoz et al., 2019). Moreover, the association patterns of SNPs and ROIs selected by MRI and PET were very similar, which indicated the ability of our model to identify modality-consistent biomarkers.

To gain more insight, we further analyzed four undocumented SNPs (rs4904901, rs7157639, rs11629428, and rs3829947) identified by MT-SCCAR with LD group information. The imaging QTs which had the strongest association with these four SNPs were singled out. Consequently, a total of eight SNP-ROI pairs were generated to validate the proposed model. These associations can also allow us to explore relationships from the microscopic molecular level to the macroscopic brain level. **Table 6** shows the Pearson correlation coefficients and p-values of eight SNP-ROI pairs. The p-values of all eight pairs were small, indicating a significant correlation within each pair. For rs4904901,it was correlated strongest with the same brain region across both imaging modalities, which suggests it is a modality-consistent association pattern. For the rest of the SNPs, the heterogeneous association patterns may have great potential to help us understand how changes in molecular level influence brain structure and metabolic.

CONCLUSION

In this paper, we proposed the MT-SCCAR model to investigate potential neuroimaging and genetic biomarkers. Compared with TSCCA and MTSCCA, the proposed model integrated genotype, multiple neuroimaging, and neuropsychological assessments into a single model to analyze multi-modal information. We tested our model on synthetic and ADNI data sets and compared its association results with those of TSCCA and MTSCCA. We found that our model demonstrated higher CCCs of 0.34 ± 0.04 (LD) and 0.39 ± 0.03 (gene) compared with the CCCs of TSCCA (0.23 ± 0.03) and MTSCCA (0.30 ± 0.06). Moreover, MT-SCCAR identified a small number of SNPs from enormous SNPs that were related to AD, wherein all of the top ten selected ROIs were AD brain risk regions. These satisfactory results show that MT-SCCAR outperforms TSCCA and MT-SCCA in detecting disease-specific biomarkers on multi-modal data.

The proposed model incorporates SNPs, neuroimaging measurements, and cognitive scores. However, there are a number of biological pathways that correlate with structural changes in the brain. Therefore, future efforts should aim to integrate data across more levels (i.e., gene expression, cell, and DNA methylation) for a more sophisticated understanding of the biological pathways leading from gene to disease.

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available but are available upon request at the following private repository: Alzheimer's Disease Neuroimaging Initiative, http://adni.loni.usc.edu/data-samples/access-data/, https://ida.loni.usc.edu/pages/access/studyData.jsp (The dataset contains the neuropsychological assessment data), and https://ida.loni.usc.edu/pages/access/geneticData.jsp (The dataset contains the genetics data). Requests to access the datasets should be directed to (Alzheimer's Disease Neuroimaging Initiative or catherine.conti@ucsf.edu). The code is available at https://github.com/ftorange/MT-SCCAR.

AUTHOR CONTRIBUTIONS

WK and FK designed the model and analyzed the results. FK prepared data and drafted the manuscript. SW and FK performed the pre-processing with imaging and genetics data. WK helped with data interpretation and manuscript drafting. All authors read and approved the final manuscript.

FUNDING

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

ACKNOWLEDGMENTS

We appreciate the Alzheimer's Disease Neuroimaging Initiative (ADNI) for contributing data.

REFERENCES

Adeli, E., Wu, G., Saghafi, B., An, L., Shi, F., and Shen, D. (2017). Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson's disease. *Sci. Rep.* 7, 41069–41069. doi: 10.1038/srep41069

Arking, D. E., Cutler, D. J., Brune, C. W., Teslovich, T. M., West, K., Ikeda, M., et al. (2008). A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.* 82, 160–164. doi: 10.1016/j.ajhg.2007.09.015

Ashburner, J., and Friston, K. (2007). "CHAPTER 7 – Voxel-based morphometry," in *Statistical Parametric Mapping*, eds K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny (London: Academic Press), 92–98.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bogdan, R., Salmeron, B. J., Carey, C. E., Agrawal, A., Calhoun, V. D., Garavan, H., et al. (2017). Imaging genetics and genomics in psychiatry: a critical review of progress and potential. *Biol. Psychiatry* 82, 165–175. doi: 10.1016/j.biopsych.2016.12.030

Botzung, A., Philippi, N., Noblet, V., Loureiro de Sousa, P., and Blanc, F. (2019). Pay attention to the basal ganglia: a volumetric study in early dementia with Lewy bodies. *Alzheimers Res. Ther.* 11, 108. doi: 10.1186/s13195-019-0568-y

Boutte, D., and Liu, J. (2010). Sparse canonical correlation analysis applied to fMRI and genetic data fusion. *Proc. IEEE Int. Conf. Bioinform. Biomed.* 2010, 422–426. doi: 10.1109/BIBM.2010.5706603

Cano, S. J., Posner, H. B., Moline, M. L., Hurt, S. W., Swartz, J., Hsu, T., et al. (2010). The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. *J. Neurol. Neurosurg. Psychiatry* 81, 1363–1368. doi: 10.1136/jnnp.2009.204008

Cervantes, S., Samaranch, L., Vidal-Taboada, J. M., Lamet, I., Bullido, M. J., Frank-Garcia, A., et al. (2011). Genetic variation in APOE cluster region and Alzheimer's disease risk. *Neurobiol Aging* 32, 2107.e7–17. doi: 10.1016/j.neurobiolaging.2011.05.023

Chen, X., and Liu, H. (2011). An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. *Stat. Biosci.* 4, 3–26. doi: 10.1007/s12561-011-9048-z

Cho, H., Kim, J. H., Kim, C., Ye, B. S., Kim, H. J., Yoon, C. W., et al. (2014). Shape changes of the basal ganglia and thalamus in Alzheimer's disease: a three-year longitudinal study. *J. Alzheimers Dis.* 40, 285–295. doi: 10.3233/JAD-132072

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Du, L., Liu, F., Liu, K., Yao, X., Risacher, S. L., Han, J., et al. (2020). Identifying diagnosis-specific genotype–phenotype associations via joint multitask sparse canonical correlation analysis and classification. *Bioinformatics* 36(Suppl.1), i371–i379. doi: 10.1093/bioinformatics/btaa434

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J., et al. (2021). Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 227–239. doi: 10.1109/TCBB.2019.2947428

Hao, X., Li, C., Du, L., Yao, X., Yan, J., Risacher, S. L., et al. (2017). Mining Outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Sci. Rep.* 7:44272. doi: 10.1038/srep44272

Horwitz, B., Rumsey, J., and Donohue, B. (1998). Functional connectivity of the angular gyrus in normal reading and dyslexia. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8939–8944.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834. doi: 10.1002/gepi.20533

Lin, E., Tsai, S.-J., Kuo, P.-H., Liu, Y.-L., Yang, A. C., and Kao, C.-F. (2017). Association and interaction effects of Alzheimer's disease-associated genes and lifestyle on cognitive aging in older adults in a Taiwanese population. *Oncotarget* 8, 24077–24087.

Liu, C., Chyr, J., Zhao, W., Xu, Y., Ji, Z., Tan, H., et al. (2018). Genome-wide association and mechanistic studies indicate that immune response contributes to Alzheimer's disease development. *Front. Genet.* 9:410. doi: 10.3389/fgene.2018.00410

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., and Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* 30, 241–255. doi: 10.1002/hbm.20508

Luck, D., Danion, J.-M., Marrer, C., Pham, B.-T., Gounot, D., and Foucher, J. (2010). The right parahippocampal gyrus contributes to the formation and maintenance of bound information in working memory. *Brain Cogn.* 72, 255–263. doi: 10.1016/j.bandc.2009.09.009

Ma, C., Zhang, Y., Li, X., Zhang, J., Chen, K., Liang, Y., et al. (2016). Is there a significant interaction effect between apolipoprotein E rs405509 T/T and epsilon4 genotypes on cognitive impairment and gray matter volume? *Eur. J. Neurol.* 23, 1415–1425. doi: 10.1111/ene.13052

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L., and Kuhl, D. E. (1997). Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Ann. Neurol.* 42, 85–94. doi: 10.1002/ana.410420114

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., et al. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* 2:e27. doi: 10.1371/journal.pgen.0020027

Moreno-Jimenez, E. P., Flor-Garcia, M., Terreros-Roncal, J., Rabano, A., Cafini, F., Pallas-Bazarra, N., et al. (2019). Adult hippocampal neurogenesis is abundant in neurologically healthy subjects and drops sharply in patients with Alzheimer's disease. *Nat. Med.* 25, 554–560. doi: 10.1038/s41591-019-0375-9

Munoz, S. S., Garner, B., and Ooi, L. (2019). Understanding the role of ApoE fragments in Alzheimer's disease. *Neurochem. Res.* 44, 1297–1305. doi: 10.1007/s11064-018-2629-1

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rasetti, R., and Weinberger, D. R. (2011). Intermediate phenotypes in psychiatric disorders. *Curr. Opin. Genet. Dev.* 21, 340–348. doi: 10.1016/j.gde.2011.02.003

Tanzi, R. E., Blacker, D., Bertram, L., McQueen, M. B., and Mullin, K. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* 39, 17–23. doi: 10.1038/ng1934

Teng, E., Becker, B. W., Woo, E., Knopman, D. S., Cummings, J. L., and Lu, P. H. (2010). Utility of the functional activities questionnaire for distinguishing mild cognitive impairment from very mild Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 24, 348–353. doi: 10.1097/WAD.0b013e3181e2fc84

Tombaugh, T. N., and McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *J. Am. Geriatr. Soc.* 40, 922–935. doi: 10.1111/j.1532-5415.1992.tb01992.x

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., et al. (2012a). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28, 229–237. doi: 10.1093/bioinformatics/btr649

Wang, H., Nie, F., Huang, H., Risacher, S. L., Saykin, A. J., Shen, L., et al. (2012b). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28, i127–i136. doi: 10.1093/bioinformatics/bts228

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164–e164. doi: 10.1093/nar/gkq603

Xiao, H., Gao, Y., Liu, L., and Li, Y. (2017). Association between polymorphisms in the promoter region of the apolipoprotein E (APOE) gene and Alzheimer's disease: a meta-analysis. *EXCLI J.* 16, 921–938. doi: 10.17179/excli2017-289

Xu, Z., Shen, X., Pan, W., and Alzheimer's Disease Neuroimaging I. (2014). Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS One* 9:e102312. doi: 10.1371/journal.pone.0102312

Yan, J., Liu, K., Lv, H., Amico, E., Risacher, S. L., Wu, Y. C., et al. (2018). "Joint exploration and mining of memory-relevant brain anatomic and connectomic patterns via a three-way association model," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018))*, (Washington, DC: IEEE), 6–9.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and Alzheimer's Disease Neuroimaging I. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Zhen, J., Huang, X., Van Halm-Lutterodt, N., Dong, S., Ma, W., Xiao, R., et al. (2017). ApoE rs429358 and rs7412 polymorphism and gender differences of serum lipid profile and cognition in aging chinese population. *Front. Aging Neurosci.* 9:248. doi: 10.3389/fnagi.2017.00248

Conflict 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ke, Kong and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.