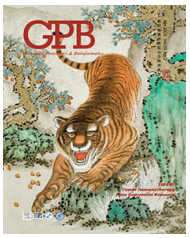
Genomics Proteomics Bioinformatics 21 (2023) 396–413

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



METHOD

inMTSCCA: An Integrated Multi-task Sparse Canonical Correlation Analysis for Multi-omic Brain Imaging Genetics

Lei Du *, Jin Zhang, Ying Zhao, Muheng Shang, Lei Guo, Junwei Han, The Alzheimer's Disease Neuroimaging Initiative †

Department of Intelligent Science and Technology, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Received 13 November 2021; revised 29 January 2023; accepted 14 March 2023
Available online 11 July 2023

Handled by Kun Huang

KEYWORDS

Brain imaging genetics;
Multi-omic endophenotype;
Cross-endophenotype association;
Genetic risk factor;
Medical image analysis

Abstract Identifying **genetic risk factors** for Alzheimer's disease (AD) is an important research topic. To date, different endophenotypes, such as imaging-derived endophenotypes and proteomic expression-derived endophenotypes, have shown the great value in uncovering risk genes compared to case–control studies. Biologically, a co-varying pattern of different omics-derived endophenotypes could result from the shared genetic basis. However, existing methods mainly focus on the effect of endophenotypes alone; the effect of **cross-endophenotype** (CEP) associations remains largely unexploited. In this study, we used both endophenotypes and their CEP associations of multi-omic data to identify genetic risk factors, and proposed two integrated multi-task sparse canonical correlation analysis (inMTSCCA) methods, *i.e.*, pairwise endophenotype correlation-guided MTSCCA (*pc*MTSCCA) and high-order endophenotype correlation-guided MTSCCA (*hoc*MTSCCA). *pc*MTSCCA employed pairwise correlations between magnetic resonance imaging (MRI)-derived, plasma-derived, and cerebrospinal fluid (CSF)-derived endophenotypes as an additional penalty. *hoc*MTSCCA used high-order correlations among these multi-omic data for regularization. To figure out genetic risk factors at individual and group levels, as well as altered endophenotypic markers, we introduced sparsity-inducing penalties for both models. We compared *pc*MTSCCA and *hoc*MTSCCA with three related methods on both simulation and real (consisting of neuroimaging data, proteomic analytes, and genetic data) datasets. The results showed that our methods obtained better or comparable canonical correlation coefficients (CCCs) and better feature subsets than benchmarks. Most importantly, the identified genetic loci and heterogeneous

* Corresponding author.
E-mail: dulei@nwpu.edu.cn (Du L).
† Consortium authors are enumerated at the end of this article.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2023.03.005

endophenotypic markers showed high relevance. Therefore, jointly using **multi-omic endophenotypes** and their CEP associations is promising to reveal genetic risk factors. The source code and manual of inMTSCCA are available at https://ngdc.cncb.ac.cn/biocode/tools/BT007330.

Introduction

Alzheimer's disease (AD) is one of the severe brain disorders and has been known as highly inheritable [1]. AD usually attack many components of the body system, *e.g.*, the brain tissue, the blood system, and the cerebrospinal fluid (CSF) [2], which could lead to many abnormal alterations. Therefore, AD patients could manifest with multiple altered endophenotypes, *e.g.*, the measurable traits at different levels of biological organization. These alterations, happening to the magnetic resonance imaging (MRI)-derived quantitative traits (QTs) and plasma-derived proteomic analytes, could co-occur based on the same genetic architecture. The co-occurrence of multiple heterogeneous endophenotypes thus could play a critical role in identifying genetic risk factors of AD [3–5].

The co-occurrence of multiple endophenotypes can result in a relatively high correlation between multiple endophenotypes, and is quantified as the cross-endophenotype (CEP) association in this study. Taking AD as an example, the apolipoprotein E (*APOE*) genotype is synchronously associated with different endophenotypes, including APOE protein levels in the CSF [6–8] and neuroimaging QTs of the reduced hippocampal volume and elevated amyloid deposition [9]. That is, both CSF levels and imaging QTs could point to the same genetic risk factors, and thus they may be correlated in the absence of AD [10]. On this account, the CEP association, in this study, exists because a genetic locus or gene is associated with more than one endophenotypes regardless of the underlying cause [3]. Since the CEP association can occur within omic data (*e.g.*, CEP associations within imaging QTs) and between multi-omic data (*e.g.*, CEP associations between imaging QTs and proteomic analytes), both intra-omic and inter-omic CEP associations could point to shared genetic factors. Therefore, the multi-omic CEP association stands a good chance of prompting the identification of genetic risk factors, which would yield new insight into the genetic architecture of AD. Pleiotropy is a similar terminology, which takes the causal effect into consideration. It refers to that a genetic locus or gene truly affects multiple endophenotypes [3]. Therefore, the pleiotropy generally focuses on identifying causal variants other than tag single nucleotide polymorphisms (SNPs). With the aim of detecting comprehensive genetic factors, the CEP association is a better choice since it has more comprehensive coverage of genetic effects. Generally, we can further distinguish between different types of pleiotropy (such as biological pleiotropy, mediated pleiotropy, and spurious pleiotropy) based on the output yielded by multi-omic CEP association studies [3,11].

In brain imaging genetics, different endophenotypes have been widely used where the associations between endophenotypes (*e.g.*, imaging QTs or proteomic analytes) and SNPs were extensively investigated [12]. Compelling evidence suggests that using carefully selected endophenotypes, *e.g.*, AD-altered imaging QTs, is able to discover novel loci that are unlikely to be revealed in case–control studies [12,13]. Roughly speaking, there are three distinct types of analytical methods, including the univariate methods, the multivariate regression methods, and the bi-multivariate correlation methods [12].

In brief, the univariate methods repeatedly analyze every pair of QT and SNP [9]. The multivariate regression methods investigate the impact of a segment of SNPs on one or multiple QTs, which can capture the group structure of multiple SNPs simultaneously [14]. The bi-multivariate correlation methods study the association between multiple QTs and multiple SNPs (multi-QTs–multi-SNPs). To the best of our knowledge, a common critical issue of these methods is that they either ignore the CEP association [9], or only involve the intra-omic CEP association [15–17]. As a result, the inter-omic CEP association remains largely unemployed. Therefore, it is essential and important to develop new methods that can employ multi-omic CEP associations including both intra-omic and inter-omic CEP associations, which could gain increased capability in identifying meaningful and reliable genetic loci [11].

Although the multi-trait genetic association study could be an effective tool for this task, they still face two difficult challenges. First, they usually use peripheral phenotypic traits which could have limited identification power compared to endophenotypic traits [3–5]. Second, they emphasize on a limited set of pre-selected traits which may ignore the associations of trait pairs, especially those seemingly distinct trait pairs being synergistically regulated by shared genetic basis [3].

In this study, to overcome the aforementioned drawbacks, we proposed two integrated multi-task sparse canonical correlation analysis (MTSCCA) methods to identify genetic factors. These two MTSCCA methods, named pairwise endophenotype correlation-guided MTSCCA (*pc*MTSCCA) and high-order endophenotype correlation-guided MTSCCA (*hoc*MTSCCA), incorporate different types of CEP associations. *pc*MTSCCA utilizes the pairwise CEP associations between MRI-derived, plasma-derived, and CSF-derived endophenotypes as an additional penalty. *hoc*MTSCCA incorporates the high-order CEP associations among MRI-derived, plasma-derived, and CSF-derived endophenotypes for regularization. To identify meaningful genetic risk factors and relevant endophenotypes of multi-omics, we used $L_{2,1}$-norm, L_1-norm, and Fused pairwise Group Lasso ($FGL_{2,1}$) penalty to conduct feature selection at different levels. The contributions of this are four-fold. First, we employed both multi-omic endophenotypes and their CEP associations, which is a better modeling strategy than existing methods. Therefore, *pc*MTSCCA and *hoc*MTSCCA possess a comprehensive and accurate ability for risk locus identification. Second, both the pairwise and high-order associations were used and verified, which could provide a diverse and useful guidance for future method development. Third, the combination of $L_{2,1}$-norm and $FGL_{2,1}$ aid to select the shared risk loci affecting multi-omic data jointly. The $FGL_{2,1}$ penalty further takes the linkage disequilibrium (LD) into consideration, which is a practical feature selection penalty. Fourth, we proposed a unified and efficient iteration optimization algorithm, and theoretically analyzed its convergence.

To evaluate *pc*MTSCCA and *hoc*MTSCCA, we compared them with three related methods, *i.e.*, one conventional sparse multiple canonical correlation analysis (SMCCA) [18] and two state-of-the-art ones, adaptive SMCCA [19], and relaxed penalized matrix decomposition canonical correlation analysis (RelPMDCCA) [20], since they are suitable for multi-omic data. We used four simulation datasets with different weight patterns and signal-to-noise (SNR) levels, and one real-world dataset containing the brain imaging QTs, proteomic analytes, and SNPs from Alzheimer's Disease Neuroimaging Initiative (ADNI) [21]. The goal is to employ multi-omic endophenotypes and their CEP associations to identify a comprehensive and meaningful subset of AD-risk loci, as well as those relevant heterogeneous endophenotypes including imaging QTs and proteomic analytes. In the simulation study, both *pc*MTSCCA and *hoc*MTSCCA obtained higher or comparable canonical correlation coefficients (CCCs) and better canonical weight profiles than all competing methods. In the real study, *pc*MTSCCA and *hoc*MTSCCA again outperformed all competitors in terms of higher CCCs and cleaner canonical weight patterns. By looking into the identified biomarkers, we found that most of the identified imaging QTs, proteomic analytes, and SNPs are related to AD. In contrast, the competitors yield too many signals with both relevant and irrelevant biomarkers reported, which could mislead the subsequent analysis. To summarize, both inMTSCCA methods are good at fusing multi-omic data to detect interesting biomarkers, which would offer a very promising new strategy for brain imaging genetics and multi-omic studies, and further deepen our understanding of the etiology and pathology of AD.

Method

In this study, we denote vectors as lowercase letters, and matrices as uppercase letters. The i-th row and j-th column of $X = (x_{ij})$ are denoted as x^i and x_j, respectively. Besides, the Euclidean norm of x is defined as $\|x\|_2 = \sqrt{\sum_i x_i^2}$. The $L_{2,1}$-norm of X is defined as $\|X\|_{2,1} = \sum_i \|x^i\|_2$, and the Frobenius norm of X is defined as $\|X\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$. Next, we briefly introduce the related SMCCA, adaptive SMCCA, and RelPMDCCA.

SMCCA

SMCCA is an extension of the two-view sparse canonical correlation analysis (SCCA) which can mine the associations among more than three views. Supposing that we face a multi-omic issue where the imaging QTs, proteomic expression markers, and SNPs are provided, SMCCA is suitable to calculate their complex associations. For ease of presentation, SNPs, plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs are associated with $X_1 \in \mathbb{R}^{n\times d}$, $X_2 \in \mathbb{R}^{n\times p_1}$, $X_3 \in \mathbb{R}^{n\times p_2}$, and $X_4 \in \mathbb{R}^{n\times q}$, respectively, where n is the number of subjects, d is the number of SNPs, p_1 is the number of plasma-derived markers, p_2 is the number of CSF-derived markers, and q is the number of imaging QTs. SMCCA can be formulated as follows:

$$\min_{u_i} \sum_{i,j=1;j\neq i}^{4} - u_i^T X_i^T X_j u_j + \sum_i R(u_i)$$

$$s.t. \|u_i\|_2^2 = 1,\ \forall i = 1,2,3,4 \quad (1)$$

where u_i is the canonical weight for each of the four omic data which indicates the contribution of each biomarker. $R(u_i)$ is the regularization term (*e.g.*, L_1-norm) to figure out a small subset of biomarkers with the highest relevance. Generally, there are tuning parameters to balance between the loss function and regularization terms. According to our previous study [22], SMCCA requires the SNP data to be associated with multiple heterogeneous endophenotypes simultaneously. This is overstrict and thus could be suboptimal, since loci affecting one omic data alone might be missed. Another obvious shortcoming is that SMCCA takes the independent assumption, *i.e.*, $X_i^T X_i = I$, which deems all features independent. This could pay for the performance degradation [20].

Adaptive SMCCA

Adaptive SMCCA improves SMCCA via adding an additional tuning parameter. Formally, adaptive SMCCA is given as follows:

$$\min_{u_i} \sum_{i,j=1;j\neq i}^{4} - \alpha_{ij} u_i^T X_i^T X_j u_j + \sum_i R(u_i)$$

$$s.t. \|u_i\|_2^2 = 1,\ \forall i = 1,2,3,4 \quad (2)$$

where α_{ij} is the parameter to balance between different sub-objectives, since the independent assumption of $X_i^T X_i = I$ will lead to biased optimization [19]. In general, adaptive SMCCA is similar to SMCCA. For one thing, it still depends on the independent assumption. For another, it is too strict to require SNPs to be associated with multi-omic data simultaneously, which is suboptimal.

RelPMDCCA

RelPMDCCA is the latest and best SMCCA to analyze multi-omic data simultaneously [20]. The definition of RelPMDCCA is similar to SMCCA, *i.e.*,

$$\min_{u_i} \sum_{i,j=1;j\neq i}^{4} - u_i^T X_i^T X_j u_j + \sum_i R(u_i)$$

$$s.t. \|X_i u_i\|_2^2 = 1,\ \forall i = 1,2,3,4 \quad (3)$$

RelPMDCCA is different to SMCCA in two aspects. First, it gets rid of the independent assumption. Second, RelPMDCCA employs the smoothly clipped absolute deviation (SCAD) penalty, which could be a more appropriate surrogate of the ideal L_0-norm than SMCCA's L_1-norm.

In summary, all aforementioned three SMCCA methods, including SMCCA, adaptive SMCCA, and RelPMDCCA, have limited capability in identifying comprehensive risk genetic loci. There are two reasons. First, all of them demand SNPs to be associated with multi-omic endophenotypes simultaneously, which is overstrict and could lead to a low recall ratio. Second, they ignore the inherent structural information

of SNPs such as LD, and thus are suboptimal for meaningful risk locus identification.

*pc*MTSCCA

The model

MTSCCA jointly learns multiple SCCA sub-objectives via multi-task learning which could identify comprehensive risk loci for multi-omic problems [22]. However, MTSCCA ignores the multi-omic CEP associations, as it only models the relationship between SNPs and multiple types of imaging QTs in parallel. To better make use of multi-omic endophenotypes and their CEP associations, we propose *pc*MTSCCA. To avoid confusion, the SNP data now are denoted as X, and the plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs are denoted as Y_1, Y_2, and Y_3, respectively. Then *pc*MTSCCA is defined as:

$$\max_{U,v_i} \sum_{i,j=1;j\neq i}^{3} \left(u_i^T X^T Y_i v_i + \Lambda_{ij}\mathrm{corr}\left(Y_i v_i, Y_j v_j\right)\right) - R(U) - \sum_i R(v_i)$$

$$s.t. \|Xu_i\|_2^2 = 1,\ \|Y_i v_i\|_2^2 = 1,\ \forall i = 1,2,3 \tag{4}$$

where Λ_{ij} is a non-negative parameter to tune the importance of CEP associations. It can be tuned by the cross-validation strategy. In addition, maximizing Equation (4) is equivalent to minimizing the following objective:

$$\min_{U,v_i} \sum_{i,j=1;j\neq i}^{3} \left(-u_i^T X^T Y_i v_i - \Lambda_{ij} v_i^T Y_i Y_j^T v_j\right) + R(U) + \sum\nolimits_i R(v_i)$$

$$s.t. \|Xu_i\|_2^2 = 1, \|Y_i v_i\|_2^2 = 1, \forall i = 1,2,3 \tag{5}$$

where U is the canonical weight for SNPs, and v_1, v_2, and v_3 are those for plasma-derived markers, CSF-derived markers, and imaging QTs, respectively. $R(U)$ and $R(v_i)$ are penalisation terms to identify relevant biomarkers. Specifically, $R(U)$ is defined as:

$$R(U) = \lambda_u \|U\|_{\mathrm{FGL}_{2,1}} + \beta \|U\|_{2,1} \tag{6}$$

where

$$\|U\|_{\mathrm{FGL}_{2,1}} = \sum\nolimits_{k=1}^{d-1} \sqrt{\|u^k\|_2^2 + \|u^{k+1}\|_2^2} \tag{7}$$

Equation (7) is the matrix form of FGL [17] and it will reduce to FGL when U degenerates to a vector. This penalty can automatically find out the proximity relationships that extensively exist among SNPs due to the LD in the human genome. λ_u is a nonnegative parameter that controls the strength of $\mathrm{FGL}_{2,1}$ penalty. Additionally, $\mathrm{L}_{2,1}$-norm helps identify whether an individual locus affects multi-omic endophenotypes jointly, thereby implicating this locus' potential pleiotropy. Hence, this hybrid penalty is a more reasonable one than that of aforementioned penalties in uncovering risk loci. The strength of this $\mathrm{L}_{2,1}$-norm is tuned by the nonnegative parameter β. In addition, we use L_1-norm for v_i to identify multi-omic endophenotypic markers, since AD will not attack all endophenotypes of them. Similarly, we use λ_i (i = 1, 2, 3) to control the sparsity of each v_i, and reasonable λ_i will help identify those relevant and important endophenotypic markers.

To sum up, *pc*MTSCCA has three advantages. Firstly, it considers both multi-omic endophenotypes and their CEP associations, and thus can be more reasonable than SMCCA, adaptive SMCCA, and RelPMDCCA. Secondly, *pc*MTSCCA has multiple SCCA tasks with each corresponding to the correlation between SNPs and one omic data. On this account, *pc*MTSCCA can make full use of each individual omic data. Thirdly, *pc*MTSCCA employs the novel $\mathrm{FGL}_{2,1}$ penalty, which takes LD into consideration while those competitors cannot.

Optimization algorithm

The *pc*MTSCCA can be equivalently rewritten as follows:

$$\min_{U,v_i} \sum\nolimits_{i,j=1;j\neq i}^{3} \left[\|Xu_i - Y_i v_i\|_2^2 + \Lambda_{ij}\|Y_i v_i - Y_j v_j\|_2^2 + R(v_i)\right] + R(U)$$

$$s.t. \|Xu_i\|_2^2 = 1, \|Y_i v_i\|_2^2 = 1, \forall i = 1,2,3 \tag{8}$$

This equation implies that there is a lower bound. According to the study by Witten and colleagues (Lemma 2.2) [23], the equality constraints can be finally satisfied by projecting the unconstrained solution onto the L_2-norm ball. Therefore, we can solve the following unconstrained problem first, *i.e.*,

$$\min_{U,v_i} \sum\nolimits_{i,j=1;j\neq i}^{3} \left[\|Xu_i - Y_i v_i\|_2^2 + \Lambda_{ij}\|Y_i v_i - Y_j v_j\|_2^2 + \lambda_i \|v_i\|_1\right] + \beta\|U\|_{2,1} + \lambda_u \sum\nolimits_{k=1}^{d-1} \sqrt{\|u^k\|_2^2 + \|u^{k+1}\|_2^2} \tag{9}$$

Similar to SMCCA, Equation (9) is multi-convex in these canonical weights, and we can solve each canonical weight alternatively with those remaining ones fixed. Without loss of generality, we solve the problem with respect to U first. When v_i is fixed, the Lagrangian function with respect to U is as follows:

$$\min_U \|XU - \mathbb{Y}\|_F^2 + \beta\|U\|_{2,1} + \lambda_u \sum\nolimits_{k=1}^{d-1} \sqrt{\|u^k\|_2^2 + \|u^{k+1}\|_2^2} \tag{10}$$

where $\mathbb{Y} = [\, Y_1 v_1 \quad Y_2 v_2 \quad Y_3 v_3 \,]$. This equation is a multi-task regression and can be easily addressed. We take its derivative with respect to U and then set it to zero, *i.e.*,

$$-X^T\mathbb{Y} + X^T X U + \beta D_u U + \lambda_u \widetilde{D}_u U = 0 \tag{11}$$

where D_u is a diagonal matrix, whose diagonal entries are $\frac{1}{2\|u^i\|_2}(i \in [1, \cdots, p])$, and $\widetilde{D}_u$ is another diagonal matrix whose diagonal entries are $\frac{1}{2\sqrt{\|u^{i-1}\|_2^2 + \|u^i\|_2^2}} + \frac{1}{2\sqrt{\|u^i\|_2^2 + \|u^{i+1}\|_2^2}}(i \in [1, \cdots, d])$ (more details are in [17]). Now we have the solution to Equation (10) as follows:

$$U = \left(X^T X + \beta D_u + \lambda_u \widetilde{D}_u\right)^{-1} X^T \mathbb{Y} \tag{12}$$

and the solution of U is attained by scaling each u_i, *i.e.*,

$$u_i = \frac{u_i}{\|Xu_i\|_2} \tag{13}$$

Once obtaining U, we can continue to solve each v_i alternatively. Similarly, considering the v_i-irrelevant terms as constants in Equation (9), we take the derivative with respect to v_i, and set it to zero, *i.e.*,

$$-Y_i^T(Xu_i - Y_i v_i) + \sum_{j=1;j\neq i}^{3} \Lambda_{ij} Y_i^T \left(Y_i v_i - Y_j v_j\right) + \lambda_i D_i v_i = 0 \tag{14}$$

where D_i is a diagonal matrix with the t-th entry being $\frac{1}{2|v_{it}|}$, and the range of t varies depending on the specific endophenotypic data as we introduced previously. Now we arrive at

$$v_i = \left(\left(1 + \sum\nolimits_j \Lambda_{ij}\right) Y_i^T Y_i + \lambda_i D_i\right)^{-1} Y_i^T \left(Xu_i + \sum\nolimits_{j=1;j\neq i}^{3} \Lambda_{ij} Y_j v_j\right) \tag{15}$$

Finally, the solution of v_i can be attained via the scaling step as follows:

$$v_i = \frac{v_i}{\|Yv_i\|_2} \tag{16}$$

Algorithm 1 contains the pseudo-code of the optimization algorithm. In this algorithm, canonical weights U and v_i are alternatively calculated till a pre-defined termination condition is satisfied. Steps 3 and 5 are easy to calculate. In the implementation, we handle Steps 4 and 6 by solving a system of linear equations, which is more efficient than calculating the matrix inverse. Therefore, the proposed algorithm could run with desirable efficiency. According to Equation (8), the *pc*MTSCCA objective has the lower bound zero. The iteration algorithm will finally attain a local optimum. In practice, the termination conditions, *i.e.*, $maxmax\left|U^{t+1} - U^t\right| \leq \varepsilon$ and $max\left|v_i^{t+1} - v_i^t\right| \leq \varepsilon$ ($\forall i \in [1, 2, 3]$), are used to assure efficiency. The tolerance error ϵ was set to 1×10^{-5} empirically.

*hoc*MTSCCA

Although *pc*MTSCCA is better than existing methods, it only cares about the pairwise correlation between heterogeneous multi-omic endophenotypes. In biomedical studies, the high-order association among multi-omic endophenotypes could also be useful, since the high-order association may implicate novel in-depth clues. To accommodate the high-order CEP association, we propose the *hoc*MTSCCA, which is defined as:

$$\max_{U,v_i} \sum_{i=1}^{3} u_i^T X^T Y_i v_i + \Lambda \mathrm{corr}(Y_1 v_1, Y_2 v_2, Y_3 v_3) - R(U) - \sum_i R(v_i)$$

$$s.t. \|Xu_i\|_2^2 = 1, \|Y_i v_i\|_2^2 = 1, \forall i = 1, 2, 3 \tag{17}$$

Further, maximizing this equation is equivalent to minimizing the following objective:

$$\min_{U,v_i} \sum_{i=1}^{3} -u_i^T X^T Y_i v_i - \Lambda C_{123} \bar{\times}_1 v_1 \bar{\times}_2 v_2 \bar{\times}_3 v_3 + R(U) + \sum_i R(v_i)$$

$$s.t. \|Xu_i\|_2^2 = 1, \|Y_i v_i\|_2^2 = 1, \forall i = 1, 2, 3 \tag{18}$$

where U and v_i are canonical weights holding the same meaning to those of *pc*MTSCCA, and so do $R(U)$ and $R(v_i)$. The second term captures the high-order canonical correlation, and Λ is used to control its contribution. According to the study by Luo and colleagues [24], C measures the covariance tensor among multi-omic data which can be calculated from

Algorithm 1 The *pc*MTSCCA algorithm

Require:

The SNP data $X \in \mathbb{R}^{n\times p}$, plasma-derived proteomic expression data $Y_1 \in \mathbb{R}^{n\times p_1}$, CSF-derived proteomic expression data $Y_2 \in \mathbb{R}^{n\times p_2}$, and imaging QT data$Y_3 \in \mathbb{R}^{n\times q}$. The parameters λ_u, β, λ_1, λ_2, λ_3, Λ_{12}, Λ_{13}, and Λ_{23}.

Ensure:

Canonical weights U and v_i.

1: Initialize $U \in \mathbb{R}^{d\times 3}$ and $v_1 \in \mathbb{R}^{p_1\times 1}$, $v_2 \in \mathbb{R}^{p_2\times 1}$, $v_3 \in \mathbb{R}^{q\times 1}$;

2: **while** not convergence **do**

3: Update D_u and $\widetilde{D}_u$ accordingly;

4: Solve U according to Equation (12), and scale each u_i according to Equation (13);

5: Update D_i accordingly;

6: Solve v_i according to Equation (15), and scale v_i according to Equation (16);

7: **end while**

8: Sorting each U and v_i in descending order based on their absolute values, and report the top K biomarkers (K is defined by the user or domain expert).

$$C_{123} = \frac{1}{n}\sum_{l=1}^{n} y_{1l} \circ y_{2l} \circ y_{3l} \tag{19}$$

where $\circ$ is the tensor product, and $\bar{\times}_k$ is the k-mode tensor-vector product.

This high-order canonical correlation quantifies the CEP associations cross multi-omic endophenotypes, and is incorporated into the *hoc*MTSCCA model. On the one hand, *hoc*MTSCCA is similar to *pc*MTSCCA. They both take into account multi-omic endophenotypes and their CEP associations. That is, *hoc*MTSCCA could outperform existing methods too. On the other hand, *hoc*MTSCCA is distinct to *pc*MTSCCA. *hoc*MTSCCA considers the high-order CEP associations, while *pc*MTSCCA emphasizes on the pairwise CEP associations. This indicates that *hoc*MTSCCA has its unique advantage in identifying risk loci. In a word, *hoc*MTSCCA and *pc*MTSCCA are in the unity of opposites, and both possess better modeling capability than existing multi-omics methods. Since *hoc*MTSCCA follows the same modeling strategy to *pc*MTSCCA, it can be solved in the same way as presented in Algorithm 1, and its convergence is also guaranteed.

Results

Experimental setup

To evaluate the proposed methods, we chose three related methods which can analyze the associations among multi-omic data and select relevant feature subsets simultaneously. They were the conventional L_1-norm penalized SMCCA [18] and two state-of-the-art ones, L_1-norm penalized adaptive SMCCA [19] and RelPMDCCA [20]. Using both conventional and state-of-the-art methods can make valid and thorough comparisons. Those other SCCA methods were not included because they cannot handle multi-omic data directly.

We employed the nested five-fold cross-validation strategy with fine-tuned parameters. In the inner loop, every combination of the candidate parameters was evaluated by cross-validation, and those yielding the best testing values were optimal parameters. Receiving the optimal parameters from the inner loop, the outer loop was used to generate final training and testing results. In this study, we set the range of candidate parameters to [0.01, 0.1, 1, 10, 100]. In some cases, this procedure might not find desirable parameters. The reason is that we usually have two goals at the same time. One is to identify a high correlation, and the other is to select relevant features. During the parameter tuning, we only focused on the correlation coefficients. To overcome this, we employed a two-stage strategy. The first stage used a large interval as we just described. In the second stage, we used a small interval to further seek the optimal parameters. In particular, if Θ was the parameter obtained from the first stage, the second stage will tune in $\Theta \pm [0.1, 0.2, \cdots, 0.5]$. In practice, this setup usually yielded reasonable parameters for both correlation and feature selection. In the experiments, all methods ran on the same setup such as the same data partition, candidate parameters, and software platform, which can make the comparison fair.

In this study, we employed the Tensor Toolbox software (https://gitlab.com/tensors/tensor_toolbox) [25] for *hoc*MTSCCA. In addition, we used feature selection ability and CCC as an indicator to evaluate the performance of the model. The CCC can be calculated as follows:

$$\mathrm{CCC} = \frac{u_i^T X_i^T Y_i v_i}{\sqrt{u_i^T X_i^T X_i u_i}\sqrt{v_i^T Y_i^T Y_i v_i}} \tag{20}$$

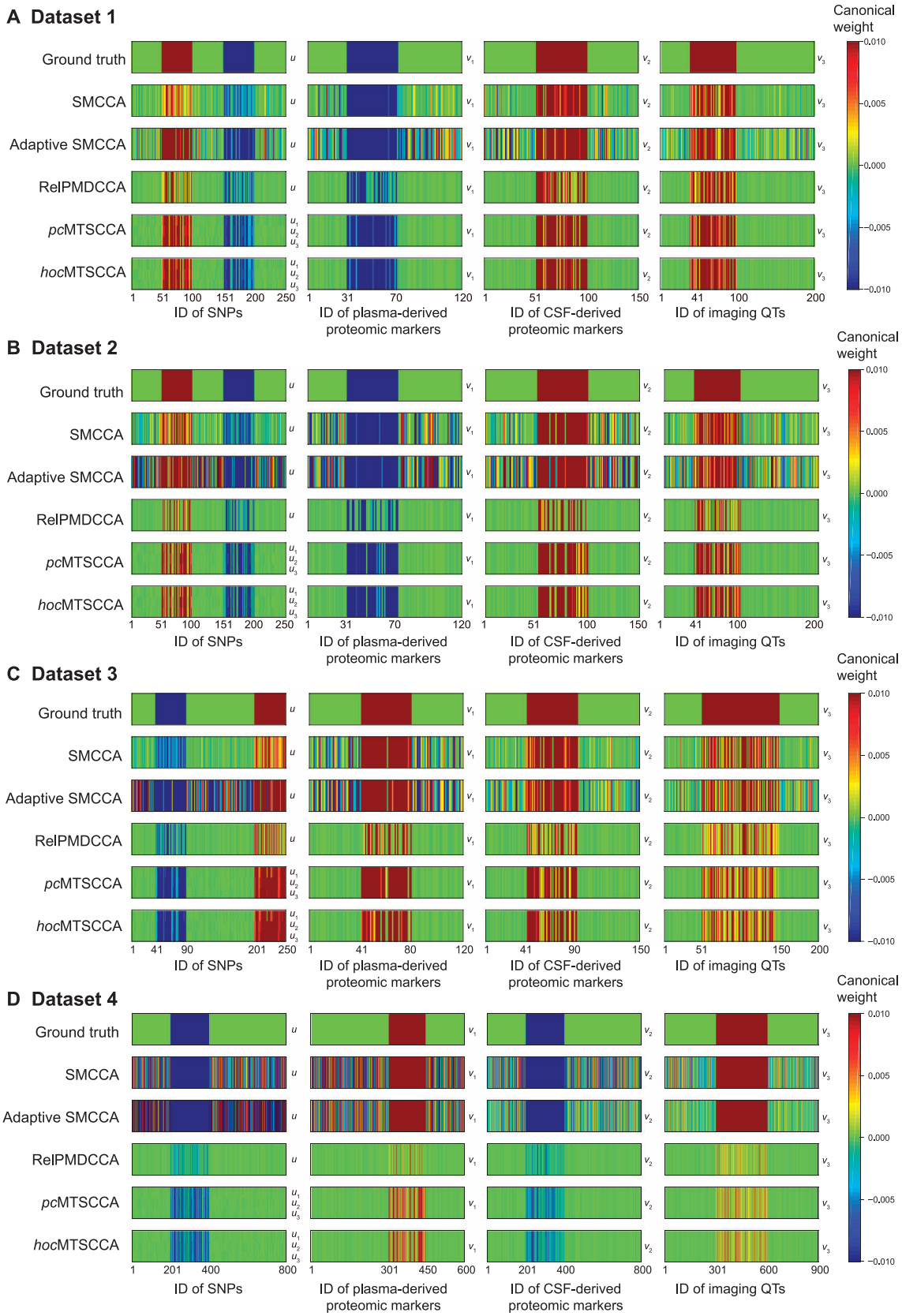
when both X_i and Y_i are normalized.

Simulation study

We generated four simulated datasets with different ground truths to ensure a thorough comparison. Each dataset had four omic data to simulate SNPs, plasma proteins, CSF proteins, and imaging QTs. Dataset 1 and Dataset 2 were generated from the same ground truth, but with different SNRs. Dataset 3 was different from the previous two datasets, with different positions and directionality of true signals. Dataset 4 had a distinct number of feature dimensionality to simulate a "small n, large p" problem. We also designed group structures to simulate the LD of genetic data for these datasets. The ground truth of each dataset was shown below and in **Figure 1** (top row in each panel).

Dataset 1: $u = (\underbrace{0,\ldots,0}_{50}, \underbrace{1,\ldots,1}_{50}, \underbrace{0,\ldots,0}_{50}, \underbrace{-1,\ldots,-1}_{50}, \underbrace{0,\ldots,0}_{50})^{\mathrm{T}}$, $v_1 = (\underbrace{0,\ldots,0}_{30}, \underbrace{-1,\ldots,-1}_{40}, \underbrace{0,\ldots,0}_{50})^{\mathrm{T}}$, $v_2 = (\underbrace{0,\ldots,0}_{50}, \underbrace{1,\ldots,1}_{50}, \underbrace{0,\ldots,0}_{50})^{\mathrm{T}}$, $v_3 = (\underbrace{0,\ldots,0}_{40}, \underbrace{2,\ldots,2}_{60}, \underbrace{0,\ldots,0}_{100})^{\mathrm{T}}$. Firstly, we set $n = 200$, $d = 250$, $p_1 = 120$, $p_2 = 150$, $q = 200$, where n denotes the number of subjects, d is the number of SNPs, p_1 is the number of plasma-derived proteomic markers, p_2 is the number of CSF-derived proteomic markers, and q is the number of imaging QTs. Then, we generated four sparse vectors, *i.e.*, $u \in \mathbb{R}^{d\times 1}$, and three sparse vectors $v_1 \in \mathbb{R}^{p_1\times 1}$, $v_2 \in \mathbb{R}^{p_2\times 1}$, $v_3 \in \mathbb{R}^{q\times 1}$. Using a latent vector $z \sim N(0, \sigma \cdot I_{n\times n})$ with $\sigma = 0.05$, we created X by $x^l \sim N(z_l u^T, \sigma \cdot \Sigma_x)$, where x^l was the l-th row of X, and

Figure 1 Canonical weights on simulated data from 20 times of five-fold cross-validation
Canonical weights of SNPs, plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs on simulated Dataset 1 (**A**), Dataset 2 (**B**), Dataset 3 (**C**), and Dataset 4 (**D**). u is the canonical weight of SNPs, v_1 is the canonical weight of plasma-derived proteomic markers, v_2 is the canonical weight of CSF-derived proteomic markers, and v_3 is the canonical weight for imaging QTs. The values in each heatmap are obtained from 20 times of trials. SMCCA, sparse multiple sparse multiple; RelPMDCCA, relaxed penalized matrix decomposition canonical correlation analysis; MTSCCA, multi-task sparse canonical correlation analysis; *pc*MTSCCA, pairwise endophenotype correlation-guided MTSCCA; *hoc*MTSCCA, high-order endophenotype correlation-guided MTSCCA; SNP, single nucleotide polymorphism; CSF, cerebrospinal fluid; QT, quantitative trait.


A Dataset 1
B Dataset 2
C Dataset 3
D Dataset 4
Ground truth
SMCCA
Adaptive SMCCA
RelPMDCCA
pcMTSCCA
hocMTSCCA
Canonical weight
0.010
0.005
0.000
−0.005
−0.010
ID of SNPs
ID of plasma-derived proteomic markers
ID of CSF-derived proteomic markers
ID of imaging QTs

$(\Sigma_x)_{i,i+1} = e^{-|u_i - u_{i+1}|}$ simulated the group structures of SNPs. Finally, we created each Y_t by $y_{i,t} \sim N(z_i v_t^T, \sigma \cdot \Sigma_{y_t})$, where Σ_{y_t} was an identity matrix. The details of the data generation procedure can be found in [26].

Dataset 2 used the same settings as Dataset 1, but with a different SNR, *i.e.*, $\sigma = 0.2$.

Dataset 3: $u = (\underbrace{0,\ldots,0}_{40}, \underbrace{-1,\ldots,-1}_{50}, \underbrace{0,\ldots,0}_{110}, \underbrace{1,\ldots,1}_{50})^{\mathrm{T}}$, $v_1 = (\underbrace{0,\ldots,0}_{40}, \underbrace{1,\ldots,1}_{40}, \underbrace{0,\ldots,0}_{40})^{\mathrm{T}}$, $v_2 = (\underbrace{0,\ldots,0}_{40}, \underbrace{2,\ldots,2}_{50}, \underbrace{0,\ldots,0}_{60})^{\mathrm{T}}$, $v_3 = (\underbrace{0,\ldots,0}_{50}, \underbrace{3,\ldots,3}_{100}, \underbrace{0,\ldots,0}_{50})^{\mathrm{T}}$, $\sigma = 0.1$.

Dataset 4: $u = (\underbrace{0,\ldots,0}_{200}, \underbrace{-1,\ldots,-1}_{200}, \underbrace{0,\ldots,0}_{400})^{\mathrm{T}}$, $v_1 = (\underbrace{0,\ldots,0}_{300}, \underbrace{1,\ldots,1}_{150}, \underbrace{0,\ldots,0}_{150})^{\mathrm{T}}$, $v_2 = (\underbrace{0,\ldots,0}_{200}, \underbrace{-2,\ldots,-2}_{200}, \underbrace{0,\ldots,0}_{400})^{\mathrm{T}}$, $v_3 = (\underbrace{0,\ldots,0}_{300}, \underbrace{1,\ldots,1}_{300}, \underbrace{0,\ldots,0}_{300})^{\mathrm{T}}$, $\sigma = 0.1$. Thus, $n = 300$, $d = 800$, $p_1 = 600$, $p_2 = 800$, $q = 900$.

We applied all methods to four simulated datasets, and employed the CCCs and canonical weight patterns to evaluate each method. A higher CCC stands for a better performance. **Table 1** presented the average CCCs from 20 times of five-fold cross-validation. For ease of presentation, we denoted the CCCs between X and Y_1, Y_2, and Y_3 as CCC1-1, CCC1-2, and CCC1-3, respectively. We observed that *pc*MTSCCA and *hoc*MTSCCA obtained higher (or comparable) CCCs than benchmark methods on both training and testing sets for most cases. In particular, Dataset 1 and Dataset 2 were generated with the same ground truth but different SNRs, and results on them demonstrated that our methods can outperform benchmarks under different noise intensities, indicating a robust performance. In addition, both proposed methods also yielded better CCCs than benchmarks on Dataset 3 and Dataset 4, where the true signals and number of features were different. In most cases on the testing CCCs, RelPMDCCA held the largest standard deviations due to its unstable performance caused by the SCAD penalty.

Besides CCCs, the heatmaps showing the feature selection are shown in Figure 1. Each benchmark had only one canonical weight u, and we repeatedly showed them for three times. This enables a clear comparison between all methods. In each panel, the average weight from 20 times of trials was shown. We observed that RelPMDCCA, *pc*MTSCCA, and *hoc*MTSCCA identified a small feature subset which was consistent to true signals. On the contrary, SMCCA and adaptive SMCCA could not identify where the true signals were. This is because their independent assumption might miss important information. In summary, these simulation results demonstrate that by eliminating the independent assumption, a method could be good at feature selection. Additionally, by fusing both multi-view data and their cross-associations, a better CCC could be obtained. We will verify this conclusion on the real ADNI data next.

Study on real neuroimaging, proteomic, and genetic data

The brain imaging data, quantification data of proteomic analytes in plasma and CSF, and genotyping data were obtained from the ADNI database. The primary goal of this initiative is to test whether serial MRI, or other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

A total of 244 non-Hispanic Caucasian participants were obtained, containing 42 healthy controls (HCs), 137 MCI patients, and 65 AD patients; the details of the participant characteristics are shown in **Table 2**. Their baseline structural MRI scans were collected and pre-processed via a widely used pipeline including the average, alignment, resample, smoothness, and normalization steps. To ensure the efficiency and increase the power, we extracted the region of interest (ROI)-level voxel-based morphometry (VBM) in the SPM software. Finally, we obtained 465 imaging QTs which were ROI measurements (gray matter density measures) spanning the whole brain based on the MarsBaR automated anatomical labeling (AAL) atlas [27]. These imaging QTs were also

Table 1 Comparison of average CCCs from 20 times of five-fold cross-validation on simulated datasets

Method	Dataset 1			Dataset 2			Dataset 3			Dataset 4		
	CCC1-1	CCC1-2	CCC1-3	CCC1-1	CCC1-2	CCC1-3	CCC1-1	CCC1-2	CCC1-3	CCC1-1	CCC1-2	CCC1-3
Training sets												
SMCCA	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.94 ± 0.01	0.94 ± 0.01	0.95 ± 0.01	**0.98 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00
Adaptive SMCCA	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.96 ± 0.00**	**0.97 ± 0.00**	**0.98 ± 0.01**	**0.98 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.98 ± 0.00	0.98 ± 0.00	**0.99 ± 0.00**
RelPMDCCA	0.96 ± 0.00	0.97 ± 0.00	0.98 ± 0.00	0.93 ± 0.01	0.93 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.97 ± 0.01
*pc*MTSCCA	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.96 ± 0.00**	0.96 ± 0.00	**0.98 ± 0.00**	**0.98 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.98 ± 0.00	**0.99 ± 0.00**	**0.99 ± 0.00**
*hoc*MTSCCA	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.95 ± 0.00	0.96 ± 0.00	**0.98 ± 0.00**	0.97 ± 0.00	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	0.98 ± 0.00	**0.99 ± 0.00**
Testing sets												
SMCCA	0.96 ± 0.00	0.97 ± 0.01	0.98 ± 0.00	0.83 ± 0.05	0.80 ± 0.04	0.87 ± 0.04	0.91 ± 0.03	0.96 ± 0.01	0.97 ± 0.01	0.94 ± 0.01	0.94 ± 0.00	0.95 ± 0.00
Adaptive SMCCA	0.96 ± 0.01	0.97 ± 0.01	0.98 ± 0.00	0.83 ± 0.03	**0.84 ± 0.04**	0.92 ± 0.03	0.91 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.94 ± 0.01	0.94 ± 0.00	0.95 ± 0.00
RelPMDCCA	0.93 ± 0.03	0.94 ± 0.02	0.96 ± 0.01	0.83 ± 0.05	0.82 ± 0.03	0.89 ± 0.02	0.88 ± 0.04	0.94 ± 0.04	0.95 ± 0.02	0.90 ± 0.02	0.89 ± 0.04	0.93 ± 0.02
*pc*MTSCCA	**0.97 ± 0.01**	**0.97 ± 0.00**	**0.99 ± 0.00**	**0.84 ± 0.03**	0.84 ± 0.05	**0.93 ± 0.02**	**0.92 ± 0.02**	**0.98 ± 0.01**	**0.98 ± 0.01**	**0.95 ± 0.02**	0.95 ± 0.02	**0.96 ± 0.01**
*hoc*MTSCCA	**0.97 ± 0.01**	0.97 ± 0.01	**0.99 ± 0.00**	0.83 ± 0.03	**0.84 ± 0.04**	**0.93 ± 0.02**	0.91 ± 0.02	0.97 ± 0.02	**0.98 ± 0.01**	**0.95 ± 0.02**	**0.95 ± 0.01**	0.95 ± 0.02

Note: The SNP data are denoted as X, and the plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs are denoted as Y_1, Y_2, and Y_3, respectively. The average CCCs (mean ± SD) between X and Y_1, Y_2, and Y_3 are denoted as CCC1-1, CCC1-2, and CCC1-3, respectively. Bold fonts represent the best results among the five methods. CCC, Canonical correlation coefficient; SMCCA, sparse multiple canonical correlation analysis; RelPMDCCA, relaxed penalized matrix decomposition canonical correlation analysis; *pc*MTSCCA, pairwise endophenotype correlation-guided multi-task sparse canonical correlation analysis; *hoc*MTSCCA, high-order endophenotype correlation-guided multi-task sparse canonical correlation analysis; SNP, single nucleotide polymorphism; CSF, cerebrospinal fluid; QT, quantitative trait; SD, standard deviation.

Table 2 Participant characteristics

Characteristic	HC	MCI	AD
Number of participants	42	137	65
Gender (male/female, %)	52.38/47.62	69.34/30.66	55.38/44.62
Handedness (right/left, %)	90.48/9.52	92.70/7.30	98.46/1.54
Age (mean ± SD, year)	75.40 ± 5.80	74.13 ± 7.22	74.75 ± 7.67
Education (mean ± SD, year)	15.88 ± 2.77	16.03 ± 2.98	15.12 ± 3.05

Note: HC, healthy control; MCI, mild cognitive impairment; AD, Alzheimer's disease.

adjusted to eliminate the effects of the baseline age, gender, handedness, and years of education.

The blood plasma and CSF samples of these 244 subjects were evaluated by the proteomic panel developed by Rules Based Medicine (RBM). After quality control (QC), we separately generated 146 plasma-derived and 83 CSF-derived proteomic markers. The genotyping data of these subjects were genotyped using the Human610-Quad or OmniExpress array platform (Illumina, San Diego, CA), and then pre-processed according to the standard QC and imputation steps. We finally obtained 1094 SNPs located in the neighborhood of AD risk gene *APOE* (boundary: ±200 kb), according to the ANNOVAR annotation. We aim to identify AD-risk loci by these heterogeneous multi-omic endophenotypes and AD-affected endophenotype markers, as well as the bi-multivariate associations between imaging QTs, proteomic markers, and SNPs.

In this real ADNI data, we compared the average CCCs of all methods (**Table 3**). The CCCs between SNPs and plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs (VBM) were denoted as SNP–Plasma, SNP–CSF, and SNP–VBM, respectively. The results showed that *pc*MTSCCA and *hoc*MTSCCA obtained higher CCCs, including both training and testing scores, for most cases than the three benchmark methods. In particular, our methods obtained much higher CCCs than competitors for SNP–Plasma and SNP–CSF CCCs. They also outperformed competitors for the SNP–VBM CCCs on testing sets. These results implies that *pc*MTSCCA and *hoc*MTSCCA yield higher values between SNPs and proteomic markers than those between SNPs and brain imaging QTs of structural MRI scans. This is very interesting since it is in agreement with the biological organization. However, all benchmark methods failed to verify this. This is also the evidence suggesting that using both multi-omic and their CEP associations can reasonably capture the relationship among multiple distinct omics data in real studies.

Evaluation of the identified biomarkers

Identification and interpretation of genetic loci

The heatmap in **Figure 2**A showed the canonical weights corresponding to the genetic data for each method. We observed that *pc*MTSCCA and *hoc*MTSCCA identified multiple relevant AD-risk loci, including the notorious common variants rs429358 (*APOE*) [28], rs56131196 (apolipoprotein C1, *APOC1*), rs4420638 (*APOC1*), rs7412 (*APOE*), rs440446 (*APOE*), and so forth. As shown in **Table 4**, all the top ten genetic loci of *pc*MTSCCA and *hoc*MTSCCA showed increased risk of AD. Although SMCCA and adaptive SMCCA could also identify some AD-related loci, both reported too many signals, indicating that they considered most of SNPs to be related to AD. Obviously, this is unconscionable. RelPMDCCA yielded a sparser result than SMCCA and adaptive SMCCA, but still denser than *pc*MTSCCA and *hoc*MTSCCA. Moreover, it missed the strongest rs429358, making its identification unconvincing. In addition, *pc*MTSCCA and *hoc*MTSCCA identified two group of SNPs, which were supported by their newly designed $FGL_{2,1}$ penalty. Specifically, SNPs with equal or very similar weight values will be reported in a same group. More details are shown in Table S1. Further analysis showed that SNPs in the same group were truly from the same LD, demonstrating that our methods are better than benchmark methods in terms of individual-level and structural-level feature selection.

Identification and interpretation of plasma-derived proteomic markers

The canonical weights showing the importance of plasma-derived proteomic markers are shown in Figure 2B. The heatmap exhibited that *pc*MTSCCA and *hoc*MTSCCA identified a small subset of plasma-based markers by assigning them a

Table 3 Comparison of average CCCs from 20 times of five-fold cross-validation on ADNI

Method	CCC for training sets			CCC for testing sets		
	SNP–Plasma	SNP–CSF	SNP–VBM	SNP–Plasma	SNP–CSF	SNP–VBM
SMCCA	0.33 ± 0.04	0.33 ± 0.04	0.29 ± 0.02	0.13 ± 0.09	0.18 ± 0.10	0.09 ± 0.06
Adaptive SMCCA	0.36 ± 0.03	0.34 ± 0.03	0.23 ± 0.02	0.17 ± 0.11	0.21 ± 0.11	0.10 ± 0.07
RelPMDCCA	0.44 ± 0.05	0.46 ± 0.03	**0.50 ± 0.04**	0.13 ± 0.12	0.11 ± 0.08	0.12 ± 0.09
*pc*MTSCCA	0.66 ± 0.04	0.44 ± 0.04	0.37 ± 0.04	**0.56 ± 0.11**	0.30 ± 0.12	**0.17 ± 0.11**
*hoc*MTSCCA	**0.66 ± 0.03**	**0.50 ± 0.04**	0.44 ± 0.04	**0.56 ± 0.11**	**0.31 ± 0.13**	0.15 ± 0.10

Note: The average CCCs (mean ± SD) between SNPs and plasma-derived proteomic markers, CSF-derived proteomic markers, and imaging QTs (VBM) are denoted as SNP–Plasma, SNP–CSF, and SNP–VBM, respectively. Bold fonts represent the best results among the five methods. VBM, voxel-based morphometry.

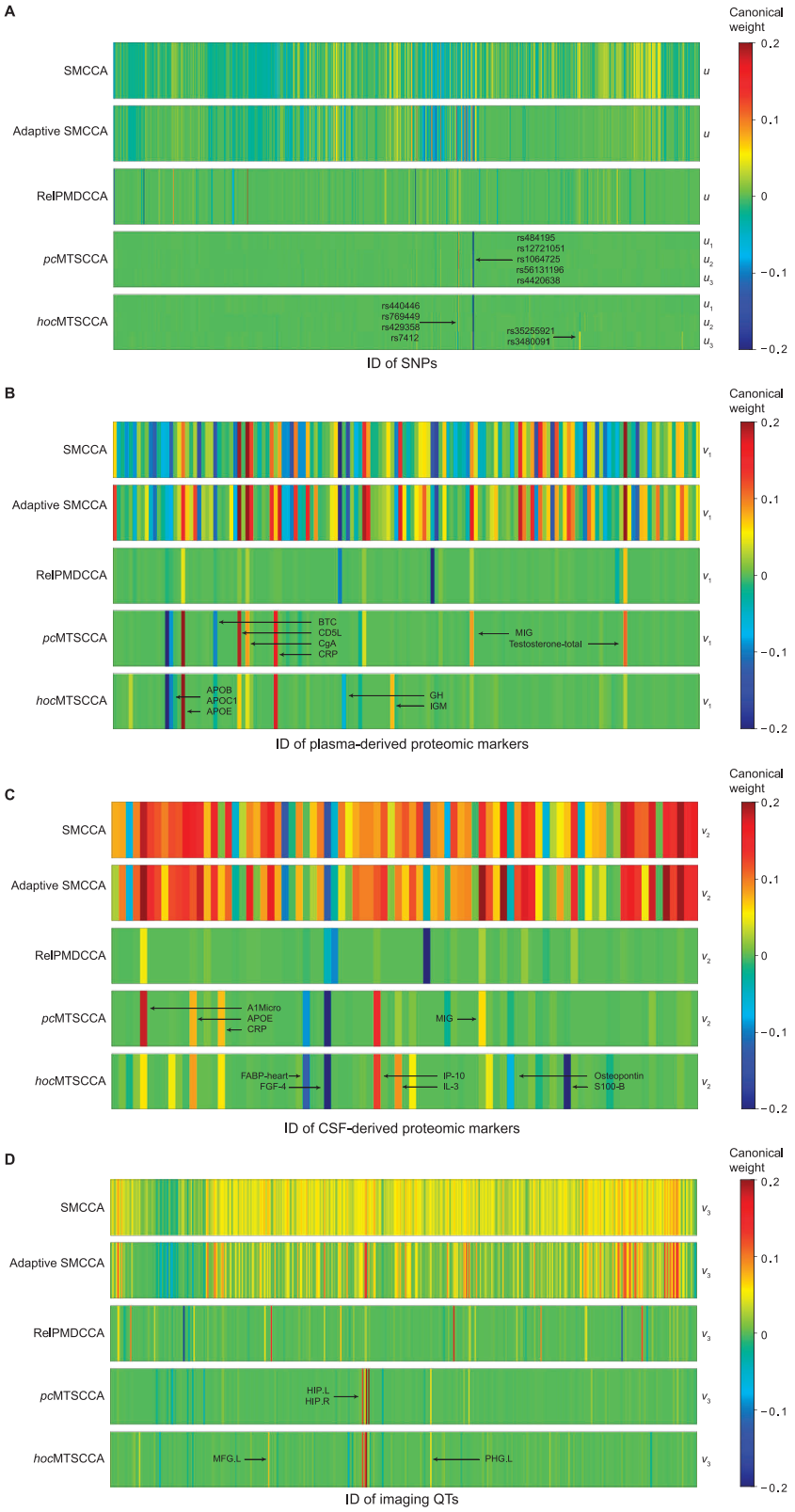


Figure 2 Canonical weights on real-word data from 20 times of five-fold cross-validation

Canonical weights of SNPs (**A**), plasma-derived proteomic markers (**B**), CSF-derived proteomic markers (**C**), and brain imaging QTs (**D**) are shown. Each heatmap shows the canonical weights corresponding to a certain method. BTC, betacellulin; CD5L, CD5 antigen-like precursor; CgA, chromogranin A; CRP, C-reactive protein; MIG, monokine induced by gamma interferon; APOB, apolipoprotein B; APOC1, apolipoprotein C1; APOE, apolipoprotein E; GH, growth hormone; IGM, immunoglobulin M; A1Micro, alpha-1-microglobulin; FABP-heart, fatty acid-binding protein-heart; FGF-4, fibroblast growth factor-4; IP-10, interferon gamma induced protein 10; IL-3, interleukin-3; S100-B, S100 calcium-binding protein B; HIP.L, hippocampus left; HIP.R, hippocampus right; PHG.L, parahippocampal gyrus left; MFG.L, middle frontal gyrus left.

Table 4 Top ten AD-related loci of each method based on average canonical weights

Method	Top ten AD-related loci
SMCCA	rs11668327, rs440446, rs73052307, rs157580, rs449647, rs11669338, rs11673139, rs17561351, rs138235833, rs41290102
Adaptive SMCCA	rs440446, rs429358, rs5117, rs483082, rs438811, rs12721051, rs56131196, rs4420638, rs11668327, rs157580
RelPMDCCA	rs11083749, rs111654618, rs111740474, rs111766460, rs77213073, rs4803776, rs41301961, rs139957871, rs35106910, rs73045691
*pc*MTSCCA	rs429358, rs56131196, rs4420638, rs7412, rs12721051, rs11668327, rs73052307, rs484195, rs157582, rs1064725
*hoc*MTSCCA	rs429358, rs56131196, rs4420638, rs7412, rs440446, rs73052307, rs12721051, rs35255921, rs34800911, rs769449

Table 5 Top ten AD-related plasma-derived proteomic markers of each method based on average canonical weights

Method	Top ten AD-related plasma-derived proteomic markers
SMCCA	Testosterone-total, CgA, FSH, CD5L, LH, PDGF-BB, TSP-1, GRO-alpha, PAI-1, RANTES
Adaptive SMCCA	Testosterone-total, CD5L, CgA, HCC-4, FSH, myoglobin, cystatin-C, IL-16, MIG, LH
RelPMDCCA	Leptin, FSH, testosterone-total, APOE, TBG, MIG, CgA, HCC-4, CD5L, APOB
*pc*MTSCCA	APOE, APOB, CD5L, CRP, testosterone-total, BTC, APOC1, MIG, CgA, HCC-4
*hoc*MTSCCA	APOE, APOB, CRP, APOC1, IGM, CD5L, GH, CgA, ACE, testosterone-total

Note: CgA, chromogranin A; FSH, follicle-stimulating hormone; CD5L, CD5 antigen-like precursor; LH, luteinizing hormone; PDGF-BB, platelet-derived growth factor BB; TSP-1, thrombospondin-1; GRO-alpha, growth-regulated alpha protein; PAI-1, plasminogen activator inhibitor 1; RANTES, T-cell-specific protein RANTES; HCC-4, chemokine CC-4; IL-16, interleukin-16; MIG, monokine induced by gamma interferon; APOE, apolipoprotein E; TBG, thyroxine-binding globulin; APOB, apolipoprotein B; CRP, C-reactive protein; BTC, betacellulin; APOC1, apolipoprotein C1; IGM, immunoglobulin M; GH, growth hormone; ACE, angiotensin-converting enzyme.

relatively higher weight value. These identified markers, such as APOE, apolipoprotein B (APOB), APOC1, C-reactive protein (CRP), CD5 antigen-like precursor (CD5L), chromogranin A (CgA), monokine induced by gamma interferon (MIG), and testosterone [29–32], have been verified to be related to AD. In contrast, SMCCA and adaptive SMCCA failed to convey useful information due to the excessively reported signals. Due to the SCAD penalty, RelPMDCCA obtained too sparse signals with many risk plasma markers missed. To make the comparison clear, we also presented the top ten identified markers in **Table 5**, and more details such as the weight values are listed in Table S2.

Identification and interpretation of CSF-derived proteomic markers

The heatmap in Figure 2C showed the canonical weights of CSF-derived proteomic markers, and the top ten selected CSF markers are listed in **Table 6** (more details are presented in Table S3). It is interesting that *pc*MTSCCA and *hoc*MTSCCA identified AD-related CSF-derived proteomic markers including APOE, CRP, FABP-heart, and MIG [33,34]. Both methods identified FGF-4 as the most relevant marker, and thus further investigation should be warranted. SMCCA and adaptive SMCCA again yielded too many markers which were very hard to interpret. Thus, they both could hardly convey useful information in real studies. RelPMDCCA committed the same mistake as it did for plasma-derived markers identification. Its over sparse results are prone to miss some meaningful markers such as APOE. Combining results obtained from both plasma-derived and CSF-derived data, we concluded that both *pc*MTSCCA and *hoc*MTSCCA performed much better than benchmarks, demonstrating their superior performance in identifying relevant proteomic expression markers.

Identification and interpretation of imaging QTs

Identifying AD-affected brain regions is also important to characterize AD. In Figure 2D, canonical weights correspond-

Table 6 Top ten AD-related CSF-derived proteomic markers of each method based on average canonical weights

Method	Top ten AD-related CSF-derived proteomic markers
SMCCA	VCAM-1, A1Micro, TRAIL-R3, TIMP-1, TM, NGAL, APOD, vWF, C3, PLGF
Adaptive SMCCA	A1Micro, NGAL, MIG, TFF3, VCAM-1, APOH, TIMP-1, PLGF, C3, HCC-4
RelPMDCCA	Leptin, FSH, A1Micro, FGF-4, MIG, SAP, PLGF, LPa, IP-10, PRL
*pc*MTSCCA	FGF-4, A1Micro, IP-10, FABP-heart, APOE, CRP, MIG, MIF, IL-3, TIMP-1
*hoc*MTSCCA	FGF-4, S100-B, IP-10, FABP-heart, IL-3, APOE, Osteopontin, A1Micro, CRP, IL-8

Note: VCAM-1, vascular cell adhesion molecule-1; A1Micro, alpha-1-microglobulin; TRAIL-R3, TNF-related apoptosis-inducing ligand receptor 3; TIMP-1, tissue inhibitor of metalloproteinase-1; TM, thrombomodulin; NGAL, neutrophil gelatinase-associated lipocalin; APOD, apolipoprotein D; vWF, von Willebrand factor; C3, complement C3; PLGF, placenta growth factor; TFF3, trefoil factor 3; APOH, apolipoprotein H; FGF-4, fibroblast growth factor-4; SAP, serum amyloid P-component; LPa, lipoprotein(a); IP-10, interferon gamma induced protein 10; PRL, prolactin; FABP-heart, fatty acid-binding protein-heart; MIF, macrophage migration inhibitory factor; IL-3, interleukin-3; S100-B, S100 calcium-binding protein B; IL-8, interleukin-8.

Table 7 Top ten AD-related brain imaging QTs of each method based on average canonical weights

Method	Top ten AD-related brain imaging QTs (ROI ID)
SMCCA	STG.R (3), ANG.R (1), HIP.L (1), MTG.R (2), STG.L (2), TPOmid.L (1)
Adaptive SMCCA	HIP.L (2), STG.R (3), STG.L (1), ANG.R (1), MTG.L (1), MTG.R (1), SPG.R (1)
RelPMDCCA	CbeCru1.L (1), IPL.R (1), MTG.L (1), MFG.L (2), TPOmid.L (1), CAL.R (1), PCUN.R (1), SFGmed.L (1), ANG.R (1)
*pc*MTSCCA	HIP.L (3), HIP.R (1), PHG.L (1), CbeCru2.R (1), Cbe8.R (1), SPG.R (1), LING.L (1), CbeCru1.R (1)
*hoc*MTSCCA	HIP.L (3), MFG.L (1), PHG.L (1), Cbe7b.R (1), ORBsup.L (1), CAU.R (1), HIP.R (1), TPOsup.L (1)

Note: The number in the parenthesis indicates the number of top ten biomarkers belonging to the same AAL brain region. ROI, region of interest; AAL, automated anatomical labeling; STG.R, superior temporal gyrus right; ANG.R, angular gyrus right; HIP.L, hippocampus left; MTG.R, middle temporal gyrus right; STG.L, superior temporal gyrus left; TPOmid.L, temporal pole: middle temporal gyrus left; MTG.L, middle temporal gyrus left; SPG.R, superior parietal gyrus right; CbeCru1.L, cerebelum crus1 left; IPL.R, inferior parietal, but supramarginal and angular gyri right; MFG.L, middle frontal gyrus left; CAL.R, calcarine fissure and surrounding cortex right; PCUN.R, precuneus right; SFGmed.L, superior frontal gyrus, medial left; HIP.R, hippocampus right; PHG.L, parahippocampal gyrus left;. CbeCru2.R, cerebelum crus2 right; Cbe8.R, cerebelum 8 right; LING.L, lingual gyrus left; CbeCru1.R, cerebelum crus1 right; Cbe7b.R, cerebelum 7b right; ORBsup.L, superior frontal gyrus, orbital part left; CAU.R, caudate nucleus right; TPOsup.L, temporal pole, superior temporal gyrus left.

ing imaging QTs are shown. We also presented the top ten selected imaging QTs in **Table 7** to make a clear comparison (more details are shown in Table S4). We observed that for this structural MRI data, *pc*MTSCCA and *hoc*MTSCCA reported that both left and right hippocampi were the primary areas being affected by AD. The left middle frontal cortical and left parahippocampal gyrus were also vulnerable. These results are consistent with previous observations that severe atrophy happens to these areas in AD patients [9]. The three benchmark methods could not provide helpful information since they reported too many irrelevant imaging QTs. That is, they considered that almost all brain areas were vulnerable to AD, which was meaningless for a clinician, since he/she cannot easily select the most relevant imaging QTs for further investigation.

To sum up, results on SNPs, two types of proteomic makers, and imaging QTs jointly demonstrate that *pc*MTSCCA and *hoc*MTSCCA can not only identify meaningful genetic loci at individual level and LD level, but also detect heterogeneous AD-related endophenotypes including plasma-derived proteomic makers, CSF-derived proteomic makers, and imaging QTs. This further demonstrates that jointly using multi-omic endophenotypes and their CEP associations could be a promising direction to accurately and comprehensively identify genetic risk factors, as well as relevant heterogeneous endophenotypes they underpin. All these conclusions confirm that our methods are powerful and practical for multi-omic fusion analysis, which can yield important clues for subsequent analysis.

Refined analysis of identified biomarkers

We next investigated the associations between SNPs and plasma-derived proteomic biomarkers, CSF-derived proteomic biomarkers, and imaging QTs. The analysis of variance (ANOVA) was also conducted to verify the effectiveness of the CEP association in identifying risk loci.

Association between selected SNPs and individual endophenotypes

The pairwise associations between SNPs and three types of endophenotypes, including plasma-derived proteomic biomarkers, CSF-derived proteomic biomarkers, and imaging QTs, are shown in **Figure 3**A and B for *pc*MTSCCA and *hoc*MTSCCA, respectively. Significant correlation scores ($P < 0.05$) were marked by "×" symbol. It was clear that most correlation scores reached the significant level, indicating the effectiveness of these markers identified by our methods. As shown in Figure 3A (top panel), the APOE expression level was significantly correlated to nine out of ten SNPs, showing its high relationship to these identified AD-risk loci. In addition, rs429358, rs56131196, and rs4420638 shared the same correlation patterns to proteomic markers herein, implicating that these three loci are located in the same group. A literature search confirmed that these three loci are in the same LD block, demonstrating the merit of the FGL penalty. The same patterns can also be observed in Figure 3B. Taken together, these results confirm that our methods can figure out meaningful pairwise associations between SNPs and endophenotypes.

Association between selected SNPs and the pairwise CEP association

To show the effectiveness of the pairwise CEP association, we conducted ANOVA to investigate the effects of plasma-derived protein concentrations, CSF-derived protein concentrations, imaging QTs, and their two-way interactions (associations) on SNPs with age, gender, years of education, and handedness as covariates. We focused on rs7412 which was missed by benchmark methods which did not consider the CEP association. The main effects of the concentrations of plasma- and CSF-derived APOE on rs7412 were all significant ($P = 5.26 \times 10^{-7}$ and $P = 4.31 \times 10^{-5}$, respectively), but that of left hippocampal volume was not ($P = 0.33$). It was interesting and meaningful that the pairwise association between the concentrations of plasma- and CSF-derived APOE and that between plasma-derived APOE concentration and left hippocampal volume were significant. These results indicate that the pairwise associations between heterogeneous endophenotypes could be beneficial to risk locus identification.

Association between selected SNPs and the high-order CEP association

The effects of high-order CEP association in assisting identification of risk loci are also of interest. The ANOVA results showed that the main effects of the concentrations of both plasma- and CSF-derived APOE on rs7412 were significant

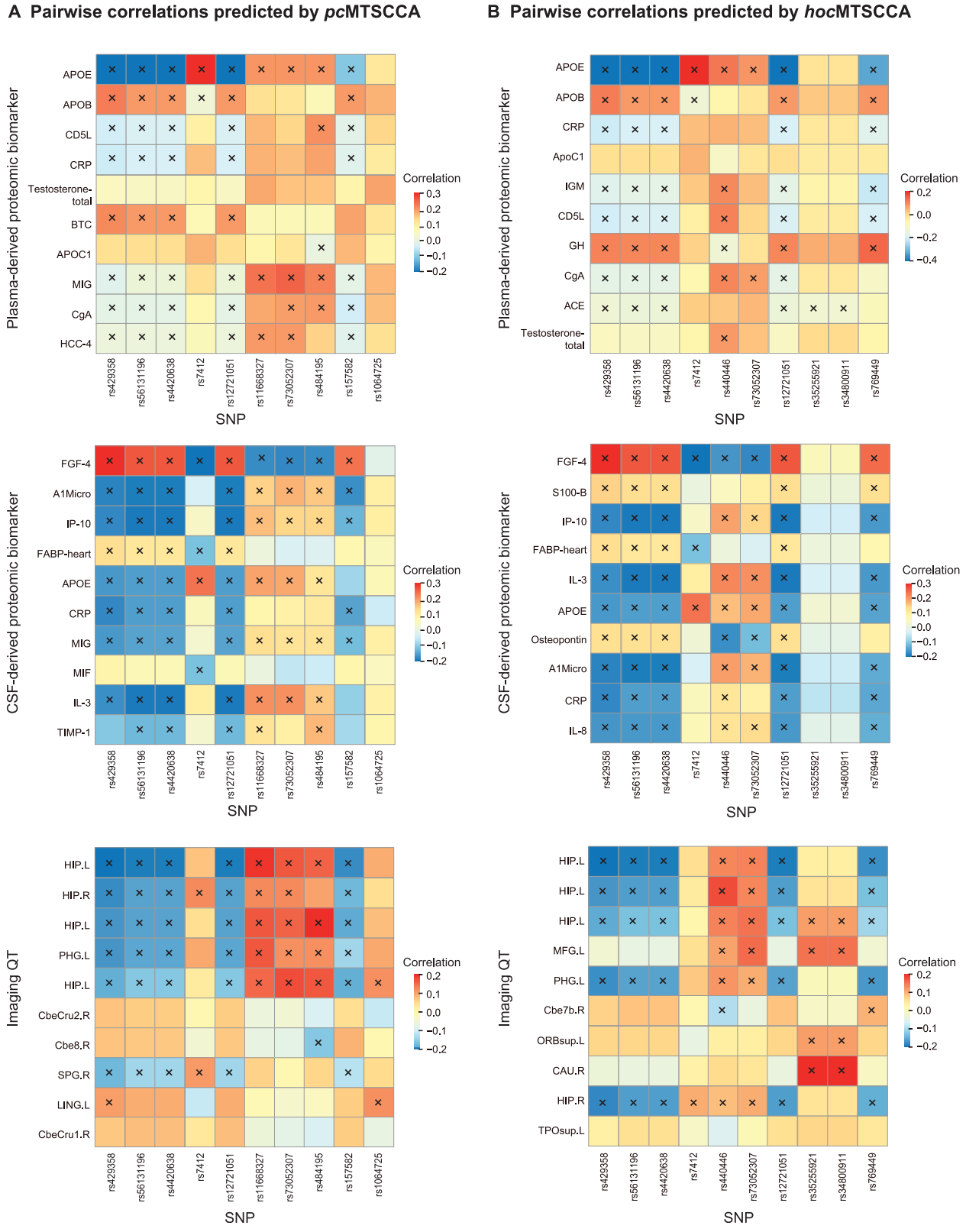


Figure 3 Pairwise correlations predicted by *pc*MTSCCA and *hoc*MTSCCA

A. Pairwise correlations predicted by *pc*MTSCCA between the identified SNPs and the plasma-derived proteomic biomarkers (top), the CSF-derived proteomic biomarkers (middle), and the imaging QTs (bottom). **B.** Pairwise correlations predicted by *hoc*MTSCCA between the identified SNPs and the plasma-derived proteomic biomarkers (top), the CSF-derived proteomic biomarkers (middle), and the imaging QTs (bottom). "×" indicates that the pairwise correlation reaches the significant level ($P < 0.05$; two-sample *t*-test). HCC-4, chemokine CC-4; MIF, macrophage migration inhibitory factor; TIMP-1, tissue inhibitor of metalloproteinases 1; ACE, angiotensin-converting enzyme; IL-8, interleukin-8; CbeCru2.R, cerebelum crus2 right; Cbe8.R, cerebelum 8 right; SPG.R, superior parietal gyrus right; LING.L, lingual gyrus left; CbeCru1.R, cerebelum crus1 right; Cbe7b.R, cerebelum 7b right; ORBsup.L, superior frontal gyrus, orbital part left; CAU.R, caudate nucleus right; TPOsup.L, temporal pole, superior temporal gyrus left.

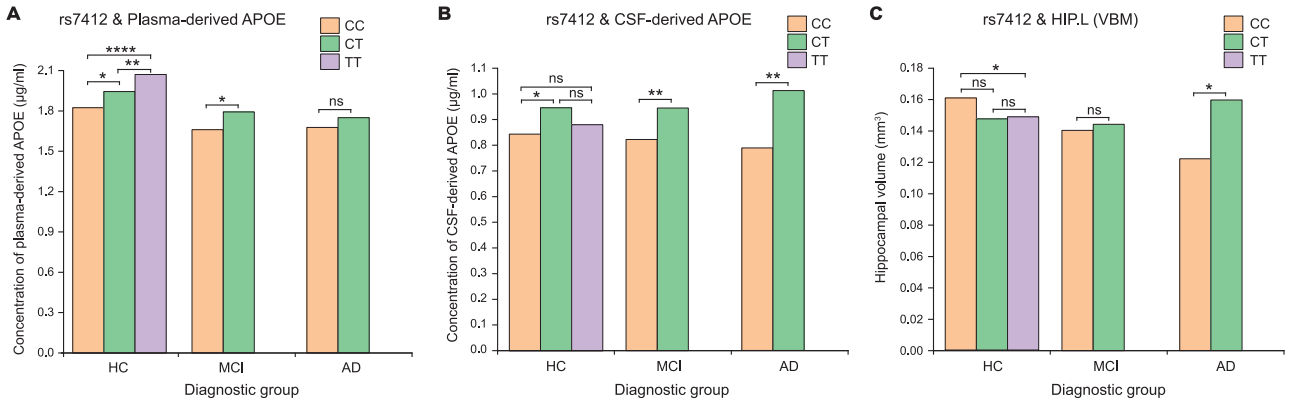


Figure 4 Pairwise comparisons for SNP and endophenotype among different diagnostic groups
A. Comparison of plasma-derived APOE concentrations for different genotypes among three diagnostic groups (HC, MCI, and AD). **B.** Comparison of CSF-derived APOE concentrations for different genotypes among three diagnostic groups (HC, MCI, and AD). **C.** Comparison of hippocampal volumes of the left hippocampus lobe for different genotypes among three groups (HC, MCI, and AD). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$; ns, not significant (two-sample *t*-test). HC, healthy control; MCI, mild cognitive impairment; AD, Alzheimer's disease; VBM, voxel-based morphometry.

($P = 5.92 \times 10^{-7}$ and $P = 4.68 \times 10^{-5}$, respectively). Surprisingly, the effect of high-order CEP associations among plasma-derived APOE concentration, CSF-derived APOE concentration, and left hippocampal volume reached the significant level ($P = 4.95 \times 10^{-4}$), even though the effect of left hippocampal volume alone was insignificant.

Then we can conclude that both pairwise and high-order CEP associations could implicate risk genetic loci, which, in all probability, is due to the shared genetic mechanism of these multiple endophenotypes. Therefore, both proposed computational methods are qualified for brain imaging genetics by taking the CEP association into consideration.

Comparisons of identified biomarkers among different diagnostic groups

We next investigated whether the selected phenotypic markers are distinctly distributed among different genotypes and diagnostic groups, such as HC, MCI, and AD. **Figure 4** showed the phenotypic distributions for plasma-derived APOE, CSF-derived APOE, and left hippocampal volume. As shown in Figure 4A, the concentrations of plasma-derived APOE in both MCI and AD groups were significantly lower than that in the HC group ($P = 4.30 \times 10^{-10}$ and $P = 1.02 \times 10^{-6}$, respectively; two-sample *t*-test), and the concentration of plasma-derived APOE in AD patients showed a decreased tendency compared to that in MCI patients ($P = 0.71$; two-sample *t*-test). In most cases, within each diagnostic group, the homozygotes of TT and heterozygotes of CT showed decreased AD risk compared to the homozygotes of CC ($P < 0.05$; two-sample *t*-test), indicating that the minor allele T could be an AD-inhibited allele, since holding this nucleotide exhibited a low risk of AD. On the contrary, carrying the major allele C might be an AD risk factor as these individuals were more vulnerable to AD attack. Similar results could also be observed in Figure 4B. As shown in Figure 4C, allele C showed increased risk to AD (*i.e.*, increased atrophy in left hippocampus), but this was not observed in homozygous CC in HC group. These results demonstrate that plasma-derived APOE concentration, CSF-derived APOE concentration, and hippocampal volume could be indicators for AD diagnosis, and the allele C of rs7412 might be an implicit factor for developing AD.

To summarize, ANOVA confirm the value of the CEP association in assisting the identification of risk loci. In addition to the endophenotype itself, its correlation(s) with other endophenotype(s) could be a good measurement in figuring out meaningful risk loci, which could increase our understanding of the genetic mechanism of AD.

Discussion

In this study, we show that both the pairwise and high-order CEP associations can improve the bi-multivariate association and feature selection. *pc*MTSCCA employs multi-omic endophenotypes and their pairwise CEP associations with due consideration given to the computation complexity, since the pairwise correlation is easy to calculate. *hoc*MTSCCA uses multi-omic endophenotypes and the high-order CEP associations, but spends more time than *pc*MTSCCA due to the tensor calculation. As a result, we suggest using *pc*MTSCCA if one intends to obtain the results quickly. On the other hand, *hoc*MTSCCA is a good alternative if one prefers the biomarker identification to time consumption.

Conclusion

AD is a highly inheritable neurodegenerative brain disorder, and many complex traits are observed in AD patients [1]. This inspires us to utilize multiple heterogeneous endophenotypes, derived from multi-omic data, to comprehensively identify genetic loci. Most existing methods cannot make use of both multi-omic endophenotypes and their CEP associations, and thus are suboptimal. Therefore, we proposed two inMTSCCA methods, *i.e.*, *pc*MTSCCA and *hoc*MTSCCA, to identify AD-related genetic factors, as well as AD-relevant endophenotypic markers. An efficient optimization algorithm was also proposed.

We compared *pc*MTSCCA and *hoc*MTSCCA with three related multi-omic methods, *i.e.*, SMCCA, adaptive SMCCA, and RelPMDCCA, on both simulated and real-world data. The results of simulated data showed that our methods obtained higher or comparable CCCs and better canonical weights under different circumstances. On the real ADNI data, both *pc*MTSCCA and *hoc*MTSCCA performed better than competitors in terms of CCCs and feature selection. The risk loci, plasma- and CSF-derived proteomic markers, and imaging QTs identified by our methods were highly related to AD. However, the three competitors either yielded too many or too little markers, which was undesirable for real biomedical studies. Taken together, our methods provide very helpful clues for further in-depth investigation and thus are powerful in identification of multi-omic heterogeneous markers. In the future, we will apply both methods to large-scale datasets, which is important for whole-genome sequence analysis.

Code availability

The inMTSCCA software tool is implemented in MATLAB (https://ww2.mathworks.cn/products/matlab.html). The code and manual have been submitted to BioCode at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (BioCode: BT007330), and are publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007330. Non-commercial use for academic, government, and non-profit institutions is permitted.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Lei Du: Conceptualization, Methodology, Writing – original draft. **Jin Zhang:** Software, Visualization, Formal analysis. **Ying Zhao:** Investigation. **Muheng Shang:** Validation. **Lei Guo:** Writing – review & editing. **Junwei Han:** Writing – review & editing. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported in part by the STI2030-Major Projects (Grant No. 2022ZD0213700), the National Natural Science Foundation of China (Grant Nos. 62136004, 61973255, and 61936007), the Natural Science Basic Research Program of Shaanxi (Grant No. 2020JM-142), and the Innovation Foundation for Doctor Dissertation at Northwestern Polytechnical University, China (Grant No. CX2023062).

Authors from ADNI

Michael W. Weiner[1], Norbert Schuff[1], Howard J. Rosen[1], Bruce L. Miller[1], David Perry[1], Arthur W. Toga[2], Lon S. Schneider[2], Sonia Pawluczyk[2], Mauricio Beccera[2], Liberty Teodoro[2], Bryan M. Spann[2], Karen Crawford[2], Scott Neu[2], Ronald Petersen[3], Bret Borowski[3], Jeff Gunter[3], Matt Senjem[3], Prashanthi Vemuri[3], David Jones[3], Kejal Kantarci[3], Chad Ward[3], Jack R. Clifford Jr.[3], Matthew Bernstein[3], Sara S. Mason[3], Colleen S. Albers[3], David Knopman[3], Kris Johnson[3], William Jagust[4], Susan Landau[4], John Q. Trojanowki[5], Leslie M. Shaw[5], Virginia Lee[5], Magdalena Korecka[5], Michal Figurski[5], Steven E. Arnold[5], Jason H. Karlawish[5], David A. Wolk[5], Christopher M. Clark[5], Laurel Beckett[6], Danielle Harvey[6], Evan Fletcher[6], Pauline Maillard[6], Charles DeCArli[6], John Olichney[6], Charles DeCarli[6], Owen Carmichael[6], Robert C. Green[7], Reisa A. Sperling[7], Keith A. Johnson[7], Gad Marshall[7], John Morris[8], Marc Raichle[8], David Holtzman[8], Nigel J. Cairns[8], Erin Franklin[8], Erin Franklin[8], Mark A. Mintun[8], Stacy Schneider[8], Angela Oliver[8], Lisa Taylor-Reinwald[8], Zaven Khachaturian[9], Greg Sorensen[10], Beau Ances[10], Maria Carroll[10], Mary L. Creech[10], Maria Carrillo[11], Lew Kuller[12], Chet Mathis[12], Oscar L. Lopez[12], Mary Ann Oakley[12], Donna M. Simpson[12], Steven Paul[13], Norman Relkin[13], Gloria Chaing[13], Michael Lin[13], Lisa Ravdin[13], Peter Davies[14], Howard Fillit[15], Franz Hefti[16], Marsel Mesulam[17], Emily Rogalski[17], Kristine Lipowski[17], Sandra Weintraub[17], Borna Bonakdarpour[17], Diana Kerwin[17], Chuang-Kuo Wu[17], Nancy Johnson[17], William Potter[18], Peter Snyder[19], Adam Schwartz[20], Tom Montine[21], Paul Aisen[22], Ronald G. Thomas[22], Michael Donohue[22], Sarah Walter[22], Devon Gessert[22], Tamie Sather[22], Gus Jiminez[22], Archana B. Balasubramanian[22], Leon Thal[22], Jennifer Mason[22], Iris Sim[22], James Brewer[22], Helen Vanderswag[22], Adam Fleisher[22], Nick Fox[23], Paul Thompson[24], Liana Apostolova[24], Kathleen Tingus[24], Ellen Woo[24], Daniel H. S. Silverman[24], Po H. Lu[24], George Bartzokis[24], Robert A. Koeppe[25], Judith L. Heidebrink[25], Joanne L. Lord[25], Norm Foster[26], Eric M. Reiman[27], Kewei Chen[27], Pierre Tariot[27], Anna Burke[27], Ann Marie Milliken[27], Nadira Trncic[27], Adam Fleisher[27], Stephanie Reeder[27], Steven Potkin[28], Adrian Preda[28], Dana Nguyen[28], Neil Buckholtz[29], John Hsiao[29], Marilyn Albert[30], Chiadi Onyike[30], Daniel D'Agostino[30], Stephanie Kielb[30], Richard Frank[31], Jeffrey Kaye[32], Joseph Quinn[32], Lisa Silbert[32], Betty Lind[32], Raina Carter[32], Sara Dolen[32], Javier Villanueva-Meyer[33], Valory Pavlik[33], Victoria Shibley[33], Munir Chowdhury[33], Susan Rountree[33], Mimi Dang[33], Rachelle S. Doody[33], Yaakov Stern[34], Lawrence S. Honig[34], Daniel Marson[35], David Geldmacher[35], Marissa Natelson Love[35], Randall Griffith[35], David Clark[35], John Brockington[35], Erik Roberson[35], Karen L. Bell[36], Hillel Grossman[36], Effie Mitsis[36], Raj C. Shah[37], Leyla deToledo-Morrell[37], Ranjan Duara[38], Maria T. Greig-Custo[38], Warren Barker[38], Martin Sadowski[39], Mohammed O. Sheikh[39], Anaztasia Ulysse[39], Mrunalini Gaikwad[39], P. Murali Doraiswamy[40], Jeffrey R. Petrella[40], Salvador BorgesNeto[40], Terence Z. Wong[40], Edward Coleman[40], Charles D. Smith[41], Greg Jicha[41], Peter Hardy[41], Partha Sinha[41], Elizabeth Oates[41], Gary Conrad[41], Anton P. Porsteinsson[42], Bonnie S. Goldstein[42], Kim Martin[42], Kelly M. Makino[42], M. Saleem Ismail[42], Connie Brand[42], Kyle Womack[43], Dana Mathews[43], Mary Quiceno[43], Allan I. Levey[44], James J. Lah[44], Janet S. Cellar[44], Jeffrey M. Burns[45], Russell H. Swerdlow[45], William M. Brooks[45], Neill R Graff-Radford[46], Francine Parfitt[46], Kim Poki-Walker[46], Christopher H. van Dyck[47], Richard E. Carson[47], Martha G. MacAvoy[47], Pradeep Varma[47], Howard Chertkow[48], Howard

Bergman[48], Chris Hosein[48], Sandra Black[49], Bojana Stefanovic[49], Curtis Caldwell[49], Ging-Yuek Robin Hsiung[50], Benita Mudge[50], Vesna Sossi[50], Howard Feldman[50], Michele Assaly[50], Elizabeth Finger[51], Stephen Pasternack[51], Irina Rachisky[51], John Rogers[51], Dick Trost[51], Andrew Kertesz[51], Charles Bernick[52], Donna Munic[52], Carl Sadowsky[53], Teresa Villena[53], Raymond Scott Turner[54], Kathleen Johnson[54], Brigid Reynolds[54], Jerome Yesavage[55], Joy L. Taylor[55], Barton Lane[55], Allyson Rosen[55], Jared Tinklenberg[55], Marwan N. Sabbagh[56], Christine M. Belden[56], Sandra A. Jacobson[56], Sherye A. Sirrel[56], Neil Kowall[57], Ronald Killiany[57], Andrew E. Budson[57], Alexander Norbash[57], Patricia Lynn Johnson[57], Thomas O. Obisesan[58], Saba Wolday[58], Joanne Allard[58], Alan Lerner[59], Paula Ogrocki[59], Curtis Tatsuoka[59], Parianne Fatica[59], Smita Kittur[60], Michael Borrie[61], Ting-Yim Lee[61], Rob Bartha[61], Sterling Johnson[61], Sanjay Asthana[61], Cynthia M. Carlsson[61], Li Shen[62], Vernice Bates[63], Horacio Capote[63], Michelle Rainka[63], Douglas W. Scharre[64], Maria Kataki[64], Brendan Kelly[64], Earl A. Zimmerman[65], Dzintra Celmins[65], Alice D. Brown[65], Godfrey D. Pearlson[66], Karen Anderson[66], Laura A. Flashman[67], Marc Seltzer[67], Mary L. Hynes[67], Robert B. Santulli[67], Karen Blank[68], Kaycee M. Sink[68], Leslie Gordineer[68], Jeff D. Williamson[68], Pradeep Garg[68], Franklin Watkins[68], Brian R. Ott[69], Geoffrey Tremont[69], Stephen Salloway[69], Paul Malloy[69], Stephen Correia[69], Lori A. Daiello[70], Jacobo Mintzer[71], Kenneth Spicer[71], David Bachman[71], Nunzio Pomara[72], Raymundo Hernando[72], Antero Sarrael[72], Susan K. Schultz[73], Karen Ekstam Smith[73], Hristina Koleva[73], Ki Won Nam[73], Hyungsub Shim[73], Amanda Smith[74], Balebail Ashok Raj[74], Kristin Fargher[74], Tatiana M. Foroud[75], Kelley Faber[75], Sungeun Kim[75], Kwangsik Nho[75], Martin R. Farlow[75], Ann Marie Hake[75], Brandy R. Matthews[75], Jared R. Brosch[75], Scott Herring[75], Andrew J. Saykin[75]

[1]*San Francisco Veterans Affairs Medical Center, University of California, San Francisco, CA 94107, USA*
[2]*Reed Neurological Research Center, UCLA School of Medicine, Los Angeles, CA 90095, USA*
[3]*Mayo Clinic Alzheimer's Disease Research Center, Rochester, MN 14607, USA*
[4]*Helen Wills Neuroscience Institute, University of California, Berkeley, CA94704, USA*
[5]*Center for Neurodegenerative Disease Researh, University of Pennsylvania, Philadelphia, PA 19019, USA*
[6]*Department of Public Health Sciences, University of California, Davis, CA 95616, USA*
[7]*Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215, USA*
[8]*Department of Pharmaceutical Chemistry, Washington University, St. Louis, MO 63110, USA*
[9]*Pitt Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 20850, USA*
[10]*Ruane Center for the Humanities, University of Pennsylvania, Philadelphia, PA 19104, USA*
[11]*Alzheimer's Association, Chicago, IL 60631, USA*
[12]*Radiology, University of Pittsburgh, Pittsburgh, PA 15213, USA*
[13]*Cornell's Medical College, Cornell University, New York, NY 14853, USA*
[14]*Albert Einstein College of Medicine, Yeshiva University, New York, NY 10461, USA*
[15]*Steering Committee of the Milken Institute's Alliance to Improve Dementia Care, AD Drug Discovery Foundation, New York, NY 10019, USA*
[16]*Acumen Pharmaceuticals, Livermore, CA 94551, USA*
[17]*Cognitive Neurology and Alzheimer's Disease Center, Northwestern University, Chicago, IL 60611, USA*
[18]*National Institute of Mental Health, Bethesda, MD 20892, USA*
[19]*Department of Neurology, Brown University, Providence, RI 02912, USA*
[20]*Tailored Therapeutics, Eli Lilly and Company, Indianapolis, IN 46285, USA*
[21]*Department of Pathology, University of Washington, Seattle, WA 98195, USA*
[22]*University of California San Diego, La Jolla, CA 92093, USA*
[23]*Department of Behavioural and Social Sciences, University of London, London WC1E 6BT, UK*
[24]*University of California, Los Angeles, Torrance, CA 90509, USA*
[25]*Department of Radiology, University of Michigan, Ann Arbor, MI 48109-2800, USA*
[26]*Department of Neurology, University of Utah, Salt Lake City, UT 84112, USA*
[27]*Banner Alzheimer's Institute, Phoenix, AZ 85006, USA*
[28]*University of California Irvine, Orange, CA 92868, USA*
[29]*National Institute on Aging, Baltimore, MD 21202, USA*
[30]*Department of Neurology, Johns Hopkins University, Baltimore, MD 21205, USA*
[31]*Richard Frank Consulting, New York, NY 13340, USA*
[32]*Oregon Health and Science University, Portland, OR 97239, USA*
[33]*Baylor College of Medicine, Houston, TX 75835, USA*
[34]*Columbia University Medical Center, New York, NY 10027, USA*
[35]*Department of Neurology, University of Alabama, Birmingham, AL 35201, USA*
[36]*Department of Neurology, Mount Sinai School of Medicine, New York, NY 75835, USA*
[37]*Rush University Medical Center, Chicago, IL 60612, USA*
[38]*Wien Center, Miami Beach, FL 33140, USA*
[39]*School of Informatics and Computing, Indiana University-Purdue University, Indianapolis, IN 46202, USA*
[40]*Departments of Radiology, Duke University Medical Center, Durham, NC 27710, USA*
[41]*Department of Neurology and Sanders-Brown Center on Aging, University of Kentucky, Lexington, KY 40292, USA*
[42]*Department of Psychiatry, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA*
[43]*Department of Neuroloy, University of Texas Southwestern Medical School, Galveston, TX 77555, USA*
[44]*Department of Neurology, Emory University, Atlanta, GA 30307, USA*
[45]*University of Kansas Medical Center, Kansas City, KS 66160, USA*
[46]*Mayo Clinic, Jacksonville, FL 28546, USA*
[47]*Yale University School of Medicine, New Haven, CT 201942, USA*
[48]*Montreal-Jewish General Hospital, Montreal, PQ H3A 2A7, Canada*
[49]*Sunnybrook Health Sciences, Toronto, ON M5S 1A5, Canada*
[50]*UBC Clinic for AD and Related Disorders, Vancouver, BC V6T 1Z2, Canada*

[51]*St. Joseph's Health Care, London, ON N6A 4H1, Canada*
[52]*Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV 89106, USA*
[53]*Premiere Research Inst (Palm Beach Neurology), West Palm Beach, FL 33413, USA*
[54]*Georgetown University Medical Center, Washington, DC 20007, USA*
[55]*Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA*
[56]*Banner Alzheimer's Institute, Phoenix, AZ 85006, USA*
[57]*Veterans Affairs Boston Healthcare System, Boston University, Boston, MA 02215, USA*
[58]*Howard University, Washington, DC 98195, USA*
[59]*Department of Neurology, Case Western Reserve University, Cleveland, OH 44106, USA*
[60]*Neurological Care of CNY, Liverpool, NY 13088, USA*
[61]*Joseph's Health Care, The University of Western Ontario, London, ON 77842, Canada*
[62]*Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine University of Pennsylvania, Philadelphia, PA 19104, USA*
[63]*Dent Neurologic Institute, Amherst, NY 14226, USA*
[64]*Department of Neurology, Ohio State University, Columbus, OH 43210, USA*
[65]*Albany Medical College, Albany, NY 12208, USA*
[66]*Hartford Hospital Olin Neuropsychiatry Research Center, Hartford, CT 06114, USA*
[67]*Dartmouth-Hitchcock Medical Center, Lebanon, NH 97355, USA*
[68]*Wake Forest School of Medicine, Winston-Salem, NC 27109, USA*
[69]*Alpert Medical School, Brown University, Providence, RI 02912, USA*
[70]*Department of Neurology, Warren Alpert Medical School of Brown University, Providence, RI 02912, USA*
[71]*Medical University South Carolina, Charleston, SC 29425, USA*
[72]*Nathan Kline Institute, Orangeburg, New York, NY 10962, USA*
[73]*University of Iowa College of Medicine, Iowa City, IA 52242, USA*
[74]*USF Health Byrd Alzheimer's Institute, University of South Florida, Tampa, FL 33613, USA*
[75]*Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2023.03.005.

ORCID

ORCID 0000-0002-6698-9814 (Lei Du)
ORCID 0000-0003-1688-5547 (Jin Zhang)
ORCID 0000-0002-3215-9986 (Ying Zhao)
ORCID 0000-0002-9227-7263 (Muheng Shang)
ORCID 0000-0003-0728-896X (Lei Guo)
ORCID 0000-0001-5545-7217 (Junwei Han)

References

[1] Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer's disease. Nat Neurosci 2020;23:311–22.
[2] Andrews SJ, Fulton-Howard B, Goate A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. Lancet Neurol 2020;19:326–35.
[3] Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet 2013;14:483–95.
[4] Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nat Methods 2015;12:755–8.
[5] Guo B, Wu B. Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. Bioinformatics 2019;35:2251–7.
[6] Riddell DR, Zhou H, Atchison K, Warwick HK, Atkinson PJ, Jefferson J, et al. Impact of apolipoprotein E (ApoE) polymorphism on brain ApoE levels. J Neurosci 2008;28:11445–53.
[7] Gupta VB, Laws SM, Villemagne VL, Ames D, Bush AI, Ellis KA, et al. Plasma apolipoprotein E and Alzheimer disease risk: the AIBL study of aging. Neurology 2011;76:1091–8.
[8] Cruchaga C, Kauwe JS, Nowotny P, Bales K, Pickering EH, Mayo K, et al. Cerebrospinal fluid APOE levels: an endopheno-type for genetic studies for Alzheimer's disease. Hum Mol Genet 2012;21:4558–71.
[9] Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. Neuroimage 2010;53:1051–63.
[10] Chiang GC, Insel PS, Tosun D, Schuff N, Truran-Sacrey D, Raptentsetsang ST, et al. Hippocampal atrophy rates and CSF biomarkers in elderly APOE2 normal subjects. Neurology 2010;75:1976–81.
[11] Tyler AL, Crawford DC, Pendergrass SA. The detection and characterization of pleiotropy: discovery, progress, and promise. Brief Bioinform 2016;17:13–22.
[12] Shen L, Thompson PM. Brain imaging genomics: integrated analysis and machine learning. Proc IEEE Inst Electr Electron Eng 2020;108:125–62.
[13] Hibar DP, Adams HH, Jahanshad N, Chauhan G, Stein JL, Hofer E, et al. Novel genetic loci associated with hippocampal volume. Nat Commun 2017;8:13624.
[14] Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. Bioinformatics 2012;28:229–37.
[15] Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med Image Anal 2014;18:891–902.
[16] Yan J, Du L, Kim S, Risacher SL, Huang H, Moore JH, et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics 2014;30:i564–71.
[17] Du L, Liu K, Yao X, Risacher SL, Han J, Saykin AJ, et al. Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. Med Image Anal 2020;61:101656.
[18] Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol 2009;8:28.
[19] Hu W, Lin D, Cao S, Liu J, Chen J, Calhoun VD, et al. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia. IEEE Trans Biomed Eng 2018;65:390–9.

[20] Rodosthenous T, Shahrezaei V, Evangelou M. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics 2020;36:4616–25.
[21] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am 2005;15:869–77.
[22] Du L, Liu K, Yao X, Risacher SL, Han J, Saykin AJ, et al. Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics. IEEE-ACM Trans Comput Biol Bioinform 2021;18:227–39.
[23] Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009;10:515–34.
[24] Luo Y, Tao D, Ramamohanarao K, Xu C, Wen Y. Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans Knowl Data Eng 2015;27:3111–24.
[25] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009;51:455–500.
[26] Chen X, Liu H. An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. Stat Biosci 2012;4:3–26.
[27] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 2002;15:273–89.
[28] Yi L, Wu T, Luo W, Zhou W, Wu J. A non-invasive, rapid method to genotype late-onset Alzheimer's disease-related apolipoprotein E gene polymorphisms. Neural Regen Res 2014;9:69–75.
[29] Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL, et al. Proteome-based plasma biomarkers for Alzheimer's disease. Brain 2006;129:3042–50.
[30] Nilsson K, Gustafson L, Hultberg B. C-reactive protein level is decreased in patients with Alzheimer's disease and related to cognitive function and survival time. Clin Biochem 2011;44:1205–8.
[31] Soares HD, Potter WZ, Pickering E, Kuhn M, Immermann FW, Shera DM, et al. Plasma biomarkers associated with the apolipoprotein E genotype and Alzheimer disease. Arch Neurol 2012;69:1310–7.
[32] Hall JR, Wiechmann AR, Cunningham RL, Johnson LA, Edwards M, Barber RC, et al. Total testosterone and neuropsychiatric symptoms in elderly men with Alzheimer's disease. Alzheimers Res Ther 2015;7:24.
[33] Mulder SD, Hack CE, van der Flier WM, Scheltens P, Blankenstein MA, Veerhuis R. Evaluation of intrathecal serum amyloid P (SAP) and C-reactive protein (CRP) synthesis in Alzheimer's disease with the use of index values. J Alzheimers Dis 2010;22:1073–9.
[34] Forlenza OV, Radanovic M, Talib LL, Aprahamian I, Diniz BS, Zetterberg H, et al. Cerebrospinal fluid biomarkers in Alzheimer's disease: diagnostic accuracy and prediction of dementia. Alzheimers Dement (Amst) 2015;1:455–63.