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

Lei Du[*, a], Jin Zhang[b], Ying Zhao[c], Muheng Shang[d], Lei Guo[e], Junwei Han[f], the Alzheimer's Disease Neuroimaging Initiative[†]

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

[*] Corresponding author(s).

E-mail: dulei@nwpu.edu.cn (Du L).

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how to apply/ADNI Acknowledgement List.pdf.

Running title: *Du et al / Integrated Multi-task Based Multi-omics Imaging Genetics*

[a]ORCID: 0000-0002-6698-9814.

[b]ORCID: 0000-0003-1688-5547.

[c]ORCID: 0000-0002-3215-9986.

[d]ORCID: 0000-0002-9227-7263.

[e]ORCID: 0000-0003-0728-896X.

[f]ORCID: 0000-0001-5545-7217.

Counts of words (abstract): 248

Counts of words (keywords): 4

Counts of words (main text): 6360

Counts of figures: 8

Counts of tables: 7

Counts of references: 34

Counts of supplementary figures: 0

Counts of supplementary tables: 4

Counts of supplementary files: 1

Counts of letters in article title: 126

Counts of letters in running title: 69

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 these different omics derived endophenotypes could result from the shared genetic basis. However, existing methods mainly focus on the effect of endophenotypes, and that of cross-endophenotype associations remains largely unexploited. In this paper, we used both endophenotypes and their cross-associations of multi-omics to identify genetic risk factors, and proposed two integrated multi-task sparse canonical correlation analysis (MTSCCA) 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 MRI-derived, plasma-derived, and cerebrospinal fluid (CSF) derived endophenotypes as an additional penalty. *hoc*MTSCCA used high-order correlations among these multi-omics for regularization. To figure out genetic risk factors at individual and group levels, as well as altered endophenotypic markers, we introduced sparsity-inducing penalties in both models. We compared *pc*MTSCCA and *hoc*MTSCCA with three related methods on both simulation data and real neuroimaging, proteomic analytes, and genetic data. The results showed that our methods obtained better or comparable canonical correlation coefficients (CCCs) and feature subsets than benchmarks. Most importantly, the identified genetic loci and heterogeneous endophenotypic markers showed high relevance. Therefore, jointly using multi-omics endophenotypes and their cross-endophenotype associations are promising to reveal genetic risk factors, and both methods are qualified for this complicated task. The source code and manual of inMTSCCA is available at: https://ngdc.cncb.ac.cn/biocode/tools/BT007330.

KEYWORDS: Brain imaging genetics; multi-omics endophenotypes; cross-endophenotype association; genetic risk factors.

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, *i.e.*, 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, *i.e.*, the measurable traits at different levels of biological organization. These alterations involve the magnetic resonance imaging (MRI) derived quantitative traits (QTs) and plasma-derived proteomic analytes, could co-occur without coincidence. The co-occurrence of multiple heterogeneous endophenotypes could play a critical role in identifying genetic risk factors of AD, given the success of multiple endophenotypic traits genetic association studies [3-5].

The co-occurrence of multiple endophenotypes could result in a relative high correlation between multiple endophenotypes and is quantified as the cross-endophenotype (CEP) association in this paper. Taking AD as an example, the apolipoprotein E(APOE) genotype is synchronously associated with different endophenotypes, including APOE protein levels in the cerebrospinal fluid (CSF) [6-8], as well as neuroimaging QTs of the reduced hippocampus volume and elevated amyloid deposition [9]. That is, both CSF levels and imaging QTs could point to the same genetic basis, and thus they may be correlated in the absence of AD [10]. On this account, the CEP association, in this paper, 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 omics data (e.g., CEP associations within imaging QTs), and between multi-omics (e.g., CEP associations between imaging QTs and proteomic analytes), both intra- and inter-omics CEP associations could point to shared genetic factors. Therefore, the multi-omics 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 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-omics 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 single nucleotide polymorphisms (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 method, the multivariate regression method and the bi-multivariate correlation method [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-omics CEP one [15-17]. As a result, the inter-omics CEP association remains largely unemployed. Therefore, it is essential and important to develop new methods that can employ multi-omics CEP associations including both intra- and inter-omics ones, which could gain increased capability in identifying meaningful and reliable genetic loci [11].

Although the multiple traits 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 paper, to overcome the above drawbacks, we propose two integrated multi-task sparse canonical correlation analysis (MTSCCA) methods to identify genetic factors. These two MTSCCA methods, called the pairwise endophenotype correlation guided MTSCCA (*pc*MTSCCA) and the 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 cerebrospinal fluid (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 use L_{21}-norm, L_1-norm and a newly designed Fused pairwise Group Lasso (FGL_{21}) penalty in conducting feature selection at different levels. The contributions of this paper are fourfold. First, we employ both multi-omics endophenotypes and their CEP associations, which is a better modelling strategy than existing methods. Therefore, both *pc*MTSCCA and *hoc*MTSCCA possess a comprehensive and accurate ability for risk loci identification. Second, both pairwise and high-order associations are used and verified, which could provide a diverse and useful guidance for future method development. Third, the combination of L_{21}-norm and FGL_{21} aid to select the shared risk loci affecting multi-omics data jointly. The FGL_{21} penalty further takes the linkage disequilibrium (LD) into consideration, which is a practical feature selection penalty. Fourth, we propose a unified and efficient iteration optimization algorithm, and theoretically analyzed its convergence.

To evaluate both *pc*MTSCCA and *hoc*MTSCCA, we compare them with three related methods, i.e. one conventional sparse multiple CCA (SMCCA) [18] and two state of-the-art ones (Adaptive SMCCA [19] and RelPMDCCA [20]), since they are suitable for multi-omics data. We use four simulation data sets with different weight patterns and signal-to-noise (SNR) levels, and one real application data set containing the brain imaging QTs, proteomic analytes and SNPs from Alzheimer's disease

neuroimaging initiative (ADNI) [21]. The goal is to employ both multi-omics 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 obtain 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 outperform all competitors in terms of higher CCCs and cleaner canonical weight patterns. By looking into the identified biomarkers, we find that most of the identified imaging QTs, proteomic analytes and SNPs are related to AD individuals. In contrast, the competitors yield too many signals with both relevant and irrelevant biomarkers reported, which could mislead the subsequent further analysis. To summarize, both integrated MTSCCA methods are good at fusing multi-omics data to detect interesting biomarkers, which would offer a very promising new strategy for brain imaging genetics and multi-omics studies, and further deepen our understanding of the etiology and pathology of AD.

Results

Experimental setup

To evaluate the proposed methods, we chose three related methods which can analyze the associations among multi-omics data and select relevant feature subsets simultaneously. They were the conventional L_1-norm penalized SMCCA [18] and two state-of-the-art ones such as L_1-norm penalized Adaptive SMCCA [19] and SCAD penalized SMCCA (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-omics data directly.

We employed the nested five-fold cross-validation strategy 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

generated final training and testing results. In this paper, 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 parameters 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 canonical correlation coefficient (CCC) as indicators to evaluate the performance of the model. The CCC can be calculated as follows

$$\mathrm{CCC}=\frac{\mathrm{u}_i^{\mathrm{T}}\mathrm{X}_i^{\mathrm{T}}\mathrm{Y}_j\mathrm{v}_j}{\sqrt{\mathrm{u}_i^{\mathrm{T}}\mathrm{X}_i^{\mathrm{T}}\mathrm{X}_j\mathrm{u}_j}\sqrt{\mathrm{v}_i^{\mathrm{T}}\mathrm{Y}_i^{\mathrm{T}}\mathrm{Y}_i\mathrm{v}_i}}$$

when both X and Y_i's are normalized.

Simulation study

Four simulated data sets were generated with different ground truths to ensure a thorough comparison. In each data set, there were four omics data including SNPs, plasma proteins, CSF proteins, and imaging QTs (denoted by four matrices X, Y_1, Y_2 and Y_3) with the same number of subjects and the different numbers of variables. Specifically, n denoted the number of subjects, d was the number of SNPs, p_1 was that of plasma-derived proteomic data, p_2 is the number of CSF-derived proteomic data, and q is the number of imaging QTs. Data 1 and Data 2 (n = 200, d = 250, p_1 = 120, p_2 = 150 and q = 200) were generated from the same ground truth, but with different signal to noise (SNR). Data 3 (n = 200, d = 250, p_1 = 120, p_2 = 150 and q = 200) was different to the previous two data sets by changing the positions and directionality of

true signals. Data 4 (n = 300, d = 800, p_1 = 600, p_2 = 800 and q = 900) had a distinct number of feature dimensionality to simulate a small-n-large-p problem. Within these data sets, we also designed group structures to simulate the LD of genetic data. We used the same data generation procedure as contained in [26], which had been widely used. The ground truth of each data was shown in **Figure 1** (top row).

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

Besides CCCs, the heatmaps showing the canonical weight patterns were 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 trials was shown. From the figure we observed that RelPMDCCA, *pc*MTSCCA and *hoc*MTSCCA identified a small feature subset which were consistent to true signals. On the contrary, SMCCA and Adaptive SMCCA cannot tell us where the true signals were. This could due to their independent assumption which might miss important information. In summary, these simulation results demonstrated 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.

Real neuroimaging, proteomic, and genetic study

The brain imaging data, quantification 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.

There were 244 non-Hispanic Caucasian participants containing 42 healthy controls (HCs), 137 MCIs and 65 Ads, and 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 (grey matter density measures) spanning the whole brain based on the MarsBaR anatomical automatic labelling (AAL) atlas [27]. These imaging QTs were also 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 Rules Based Medicine, Inc. (RBM) proteomic panel. After quality control (QC), we separately generated 146 plasma-derived proteomic markers and 83 CSF-derived ones. These subjects' genotyping data were genotyped using the Human 610-Quad or OmniExpress Array platform (Illumina, Inc., San Diego, CA, USA). These genotype data were then pre-processed according to the standard quality control (QC) and imputation steps. We finally contained 1,094 SNPs located in the neighbour of AD risk gene *APOE* (boundary: ± 200kb) according to the ANNOVAR annotation. We aim to identify AD-risk loci by these heterogeneous multi-omics endophenotypes,

AD-affected endophenotype markers, as well as the bi-multivariate association between imaging QTs, proteomic markers and SNPs.

In this real ADNI data, we presented the CCCs of all methods in **Table 3** for comparison. We denoted the CCC between SNPs and plasma-derived proteomic markers as SNP-Plasma for simplicity. Those remaining CCCs can be represented similarly. We observed that both *pc*MTSCCA and *hoc*MTSCCA obtained higher CCCs, including both training and testing scores, for most cases than comparison 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. This implies that both *pc*MTSCCA and *hoc*MTSCCA yielded 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. But all comparison methods failed to verify this. This is also the evidence suggesting that using both multi-omics and their CEP associations can reasonably capture the relationship among multiple distinct omics data in real studies.

Identification and interpretation of genetic loci

The heatmap in **Figure 2** showed the canonical weights corresponding to the genetic data for each method. We clearly observed that *pc*MTSCCA and *hoc*MTSCCA identified multiple relevant AD-risk loci, including the notorious common variant rs429358 (*APOE*) [28], rs56131196 (*APOC1*), rs4420638 (*APOC1*), rs7412 (*APOE*), rs440446 (*APOE*) and so forth. In **Table 4**, all the top ten genetic loci of *pc*MTSCCA and *hoc*MTSCCA have been shown to increase the risk of AD. Although SMCCA and Adaptive SMCCA could also identify some AD-related loci, they reported too many signals which indicates that they consider most of SNPs are related to AD. Obviously, this is unconscionable. RelPMDCCA yielded a sparser result than SMCCA and Adaptive SMCCA but still denser than ours. And, it missed the strongest rs429358 which makes its identification unconvincing. In addition, *pc*MTSCCA and *hoc*MTSCCA clearly identified two group of SNPs, which was 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 can be observed in **Table S1** in the supplementary. After investigation, we found that SNPs in the same group were truly from the same LD, which demonstrated that our methods were better than competitors in terms of both 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 were shown in **Figure 3**. The figure clearly exhibited that *pc*MTSCCA and *hoc*MTSCCA identified a small subset of all plasma-based markers by assigning them a relative higher weight value. These identified markers, such as ApoE, ApoB, ApoC-I, CRP, CD5L, CgA, MIG, Testosterone and so on [29-32], have been verified to be related to AD. In contrast, SMCCA and Adaptive SMCCA failed to convey useful information for 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 the table with more details such as the weight values was contained in the supplementary (**Table S2**). Next, we will show the results in terms of CSF-derived proteomic markers.

Identification and interpretation of CSF-derived proteomic markers

Figure 4 showed heatmap regarding canonical weights of CSF-derived proteomic markers, and **Table 6** contained the top ten selected CSF markers (more details were presented in **Table S3**). It is interesting that both our methods successfully identified AD-related CSF-derived proteomic markers including ApoE, CRP, FABP-heart, MIG and so forth [33, 34]. Both methods identified the FGF-4 as the most relevant and 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- 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 5**, canonical weights corresponding imaging QTs were shown. We also presented the top ten selected imaging QTs in **Table 7** (see **Table S4** in the supplementary for more details) to make a clear comparison. We observed that for this structural MRI data, *pc*MTSCCA and *hoc*MTSCCA reported that both left and right hippocampus were the most areas being affected by AD. The left middle frontal cortical and left parahippocampal gyrus were also vulnerable. This was in agreement to previous studies that severe atrophy happened to these areas in AD patients. The benchmarks 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 demonstrated 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- and CSF-derived proteomic makers and imaging QTs. This further demonstrated that jointly using multi-omics 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 confirmed that our methods are powerful and practical for multi-omics fusion analysis, which can yield important clues for subsequent analysis.

Refined analysis of identified biomarkers

Following up on the identified multi-omics biomarkers yielded by proposed methods, we investigate the association between SNPs and plasma-derived proteomic expression biomarkers, CSF-derived ones, as well as imaging QTs. The ANOVA was also conducted to verify the effectiveness of the CEP association in identifying risk loci.

Association between selected SNPs and the individual endophenotype

The pairwise associations between SNPs and three separate types of endophenotypes, including the plasma-derived proteomic expression biomarkers, CSF-derived ones and imaging QTs were shown in **Figure 6** and **Figure 7** for *pc*MTSCCA and *hoc*MTSCCA respectively. Those significant values ($p < 0.05$) were marked by ”×” symbols. It was clear that most association scores reached the significant level, indicating the effectiveness of these markers identified by our methods. In **Figure 6**, the subfigure (A) showed that the apolipoprotein E (ApoE) concentration 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, which implicated that these three loci were located in the same group. A literature search confirmed that these three loci were in the same LD block, which demonstrated the merit of the FGL penalty. The same patterns can be observed in **Figure 7** too. Both figures showed that our methods could figure out meaningful pairwise association between SNPs and endophenotypes.

Association between selected SNPs and the pairwise cross-endophenotype association

To show the effectiveness of the pairwise cross-endophenotypes association, we conducted ANOVA to investigate the effects of two (plasma- and CSF-derived) protein expression levels, imaging QTs and their two-way interactions (associations) on SNPs with age, gender, years of education and handedness as covariates. We focused on investigating rs7412 which was missed by those comparison methods without consideration of the CEP association. The main effects of the concentration of plasma-APOE and CSF-APOE were all significant ($p = 5.26\times10^{-7}$ and $p = 4.31\times10^{-5}$), but that of hippocampal volume was not ($p = 0.33$). The interesting and meaningful things are that the pairwise association between plasma-APOE and CSF-APOE concentrations, and that between plasma-APOE concentrations and hippocampal volume were significant. This indicated that the pairwise association between heterogeneous endophenotypes could be beneficial to risk loci identification.

Association between selected SNPs and the high-order cross-endophenotype association

We were also interested in the effects of high-order CEP association in assisting the identification of risk loci. The age, gender, years of education and handedness were included as covariates. The ANOVA results showed that main effects of both plasma-APOE and CSF-APOE were significant with $p = 5.92\times10^{-7}$ and $p = 4.68\times10^{-5}$ respectively. Surprisingly, the effect of high-order CEP among plasma-APOE, CSF-APOE and hippocampal volume reached the significant level ($p = 4.95\times10^{-4}$) even though the hippocampal volume alone was insignificant.

Combined results here and that of the last subsection, we concluded 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

In this subsection, we investigated that if the selected phenotypic markers were distinctly distributed among different genotypes and diagnostic groups such as HC, MCI and AD. **Figure 8** showed the phenotypic distributions for plasma (APOE), CSF (APOE) and hippocampal volume. In sub-figure (A), decreased plasma's (APOE) concentration level was observed in both MCI and AD groups, and the concentration level of ADs was lower than that of MCIs. Within each diagnostic group, the homozygote TT and heterozygote CT showed decreased Alzheimer's risk, 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 participants were more vulnerable to AD attack. Similar conclusions could also be observed in sub-figures (B) and (C). The brain atrophy in sub-figure (C) showed that the nucleotide C showed increased risk to ADs, but this could not be observed in homozygous CC in HC group. These results demonstrated that the concentration level of plasma (APOE), CSF (APOE) 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, the statistical analysis confirmed the value of the CEP association in assisting the identification of risk loci. In addition to the endophenotype itself, their correlation could be a good measurement in figuring our meaningful risk loci, which could increase our understanding of the genetic mechanism of AD.

Discussion

We have showed that both the pairwise CEP association and high-order one can improve the bi-multivariate association and feature selection. *pc*MTSCCA employs multi-omics endophenotypes and their pairwise CEP association with due consideration given to the computation complexity since the pairwise correlation is easy to calculate. *hoc*MTSCCA uses multi-omics endophenotypes and the high-order CEP association, 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.

Materials and methods

In this article, 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_{21}-norm of X is defined as $\|X\|_{2,1} = \sum_i \|x^i\|_2$, and the Frobenius norm is defined as $\|X\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$

Background

In this subsection, we first briefly introduced the related sparse multiple canonical correlation analysis (SMCCA), the Adaptive SMCCA and the Relaxed PMDCCA (RelPMDCCA).

Sparse multiple canonical correlation analysis (SMCCA)

SMCCA is an extension of the two-view SCCA which can mine the associations among more than three views. Suppose we face a multi-omics 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 ones and imaging QTs are associated with $\mathrm{X}_1 \in R^{n\times d}$, $\mathrm{X}_2 \in R^{n\times p_1}$, $\mathrm{X}_3 \in R^{n\times p_2}$ and $\mathrm{X}_4 \in R^{n\times q}$ 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. Now the SMCCA can be formulated as follows

$$\begin{aligned} &\min_{\mathrm{u}_i} \sum_{i,j=1;j\neq i}^{4} -\mathrm{u}_i^{\mathrm{T}}\mathrm{X}_i^{\mathrm{T}}\mathrm{X}_j\mathrm{u}_j + \sum_i R(\mathrm{u}_i) \\ &s.t. \|\mathrm{u}_i\|_2^2 = 1, \forall i = 1,2,3,4. \end{aligned} \tag{1}$$

Here u_i's are the canonical weights for four omics data which indicates the contribution of each biomarker. $\mathrm{R}(\mathrm{u}_i)$'s are the regularization terms (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 [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 omics data alone might be missed. Another obvious shortcoming is that SMCCA takes the independent assumption, *i.e.* $\mathrm{X}_i^T\mathrm{X}_i = \mathrm{I}$, which deems all features are independent. This could pay for the performance degradation [20].

Adaptive SMCCA

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

$$\begin{aligned} &\min_{\mathrm{u}_i} \sum_{i,j=1;j\neq i}^{4} -\alpha_{ij}\mathrm{u}_i^{\mathrm{T}}\mathrm{X}_i^{\mathrm{T}}\mathrm{X}_j\mathrm{u}_j + \sum_i R(\mathrm{u}_i) \\ &s.t. \|\mathrm{u}_i\|_2^2 = 1, \forall i = 1,2,3,4. \end{aligned} \tag{2}$$

$\alpha_{ij}'s$ are parameters to balance between different sub-objectives since the independent assumption of $\mathrm{X}_i^T\mathrm{X}_i = \mathrm{I}$ 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-omics data simultaneously, which is suboptimal.

RelPMDCCA

By now, RelPMDCCA is the latest and best SMCCA which has been proposed to analyze multi-omics data simultaneously [20]. The definition of RelPMDCCA is similar to SMCCA, i.e.

$$\begin{aligned} &\min_{\mathrm{u}_i} \sum_{i,j=1;j\neq i}^{4} -\mathrm{u}_i^{\mathrm{T}}\mathrm{X}_i^{\mathrm{T}}\mathrm{X}_j\mathrm{u}_j + \sum_i R(\mathrm{u}_i) \\ &s.t. \left\|\mathrm{X}_i\mathrm{u}_i\right\|_2^2 = 1, \forall i = 1,2,3,4. \end{aligned} \tag{3}$$

However, 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 above 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-omics endophenotypes simultaneously, which is overstrict and could lead to a low recall ratio. Second, they ignore the inherent structural information of SNPs such as the LD, and thus are suboptimal for meaningful risk loci identification.

***pc*MTSCCA**

The Model

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

$$\max_{\mathrm{U},\mathrm{v}_i} \sum_{i,j=1;j\neq i}^{3} \left(\mathrm{u}_i^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}_i\mathrm{v}_i + \Lambda_{ij}\mathrm{corr}\left(\mathrm{Y}_i\mathrm{v}_i, \mathrm{Y}_j\mathrm{v}_j\right)\right) - R\left(\mathrm{U}\right) - \sum_i R\left(\mathrm{v}_i\right) \quad s.t. \left\|\mathrm{X}\mathrm{u}_i\right\|_2^2 = 1, \left\|\mathrm{Y}_i\mathrm{v}_i\right\|_2^2 = 1, \forall i = 1,2,3. \tag{4}$$

where Λ_{ij}'s are nonnegative parameters to tune the importance of CEP associations. They can be tuned by the cross-validation strategy. In addition, maximizing Eq. (4) is equivalent to minimizing the following objective

$$\min_{\mathrm{U},\mathrm{v}_i} \sum_{i,j=1;j\neq i}^{3} \left(-\mathrm{u}_i^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}_i\mathrm{v}_i - \Lambda_{ij}\mathrm{v}_i^{\mathrm{T}}\mathrm{Y}_i\mathrm{Y}_j^{\mathrm{T}}\mathrm{v}_j\right) + R\left(\mathrm{U}\right) + \sum_i R\left(\mathrm{v}_i\right) \quad s.t. \left\|\mathrm{X}\mathrm{u}_i\right\|_2^2 = 1, \left\|\mathrm{Y}_i\mathrm{v}_i\right\|_2^2 = 1, \forall i = 1,2,3. \tag{5}$$

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

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

where

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

This penalty 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-omics endophenotypes jointly, thereby implicating this locus' potential pleiotropy. Hence this hybrid penalty is a more reasonable one than that of aforementioned methods 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's for those multi-omics 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's will help identify those relevant and important endophenotypic markers.

To sum up, *pc*MTSCCA has three advantages. Firstly, it considers both multi-omics 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 omics data. On this account, *pc*MTSCCA can make full use of each individual omics data. Thirdly, *pc*MTSCCA employs the novel $FGL_{2,1}$ penalty which takes the LD into consideration while those competitors cannot.

Optimization algorithm

The *pc*MTSCCA can be equivalently rewritten as

$$\begin{aligned} &\min_{\mathrm{U},\mathrm{v}_i} \sum_{i,j=1;j\neq i}^{3} \left[\left\| \mathrm{X}\mathrm{u}_i - \mathrm{Y}_i\mathrm{v}_i \right\|_2^2 + \Lambda_{ij} \left\| \mathrm{Y}_i\mathrm{v}_i - \mathrm{Y}_j\mathrm{v}_j \right\|_2^2 + R\left(\mathrm{v}_i\right) \right] + R\left(\mathrm{U}\right) \\ &s.t. \left\| \mathrm{X}\mathrm{u}_i \right\|_2^2 = 1, \left\| \mathrm{Y}_i\mathrm{v}_i \right\|_2^2 = 1, \forall i = 1,2,3. \end{aligned} \tag{8}$$

This new formula implies that there is a lower bound. According to [23] (Lemma 2.2), 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.*

$$\begin{aligned} &\min_{\mathrm{U},\mathrm{v}_i} \sum_{i,j=1;j\neq i}^{3} \left[\left\| \mathrm{X}\mathrm{u}_i - \mathrm{Y}_i\mathrm{v}_i \right\|_2^2 + \Lambda_{ij} \left\| \mathrm{Y}_i\mathrm{v}_i - \mathrm{Y}_j\mathrm{v}_j \right\|_2^2 + \lambda_i \left\| \mathrm{v}_i \right\|_1 \right] \\ &+ \beta \left\| \mathrm{U} \right\|_{2,1} + \lambda_u \sum_{k=1}^{d-1} \sqrt{\left\| \mathrm{u}^k \right\|_2^2 + \left\| \mathrm{u}^{k+1} \right\|_2^2} \end{aligned} \tag{9}$$

Similar to SMCCA, Eq. (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's are fixed, the Lagrangian with respect to U is

$$\min_{\mathrm{U}} \left\| \mathrm{X}\mathrm{U} - \mathbb{Y} \right\|_{\mathrm{F}}^2 + \beta \left\| \mathrm{U} \right\|_{2,1} + \lambda_u \sum_{k=1}^{d-1} \sqrt{\left\| \mathrm{u}^k \right\|_2^2 + \left\| \mathrm{u}^{k+1} \right\|_2^2} \tag{10}$$

where $\mathbb{Y} = [\mathbf{Y}_1\mathbf{v}_1 \quad \mathbf{Y}_2\mathbf{v}_2 \quad \mathbf{Y}_3\mathbf{v}_3]$. This equation is a multitask regression and can be easily addressed. We take its derivative with respect to U and then set it to zero, i.e.

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

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

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

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

$$\mathrm{u}_i=\frac{\mathrm{u}_i}{\left\|\mathrm{X}\mathrm{u}_i\right\|_2} \tag{13}$$

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

$$-\mathrm{Y}_i^{\mathrm{T}}\left(\mathrm{X}\mathrm{u}_i-\mathrm{Y}_i\mathrm{v}_i\right)+\sum_{j=1;\,j\neq i}^{3}\Lambda_{ij}\mathrm{Y}_i^{\mathrm{T}}\left(\mathrm{Y}_i\mathrm{v}_i-\mathrm{Y}_j\mathrm{v}_j\right)+\lambda_i\mathrm{D}_i\mathrm{v}_i=0. \tag{14}$$

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

$$\mathrm{v}_i=\left(\left(1+\sum_j\Lambda_{ij}\right)\mathrm{Y}_i^{\mathrm{T}}\mathrm{Y}_i+\lambda_i\mathrm{D}_i\right)^{-1}\mathrm{Y}_i^{\mathrm{T}}\left(\mathrm{X}\mathrm{u}_i+\sum_{j=1;\,j\neq i}^{3}\Lambda_{ij}\mathrm{Y}_j\mathrm{v}_j\right) \tag{15}$$

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

$$\mathrm{v}_i=\frac{\mathrm{v}_i}{\left\|\mathrm{Y}\mathrm{v}_i\right\|_2} \tag{16}$$

Algorithm 1 contains the pseudo-code of the optimization algorithm. In this algorithm, canonical weights U and v_i's 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 Eq. (8), the *hoc*MTSCCA objective has the lower

bound zero. The iteration algorithm will finally attain a local optimum. In practice, the termination conditions, i.e., $\max\max\left|\mathbf{U}^{t+1}-\mathbf{U}^{t}\right|\le\varepsilon$ $\max\left|\mathbf{v}_i^{t+1}-\mathbf{v}_i^{t}\right|\le\varepsilon$ ($\forall i\in[1,2,3]$), are used to assure efficiency. The tolerance error ϵ was set to 10^{-5} empirically.

Algorithm 1: The *pc*MTSCCA algorithm

Requrire:

The SNP data $\mathbf{X}\in\mathbb{R}^{n\times p}$,plasma-derived proteomic expression data $\mathbf{Y}_1\in\mathbb{R}^{n\times p_1}$,CSF-derived proteomic expression data $\mathbf{Y}_2\in\mathbb{R}^{n\times p_2}$, and imaging QT data $\mathbf{Y}_3\in\mathbb{R}^{n\times q}$ .The parameters $\lambda_u,\beta,\ \lambda_1,\ \lambda_2,\ \lambda_3,\Lambda_{12},\Lambda_{13},\Lambda_{23}$.

Ensure:

Canonical weights U and v_i's.

1: Initialize $\mathbf{U}\in\mathbb{R}^{d\times 3}$ and $\mathbf{v}_1\in\mathbb{R}^{p_1\times 1}$, $\mathbf{v}_2\in\mathbb{R}^{p_2\times 1}$, $\mathbf{v}_3\in\mathbb{R}^{q\times 1}$;

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

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

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

5: Update D_i accordingly;

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

7: **end while**

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

Extension to high-order endophenotype correlation guided MTSCCA (*hoc*MTSCCA)

Although *pc*MTSCCA is better than existing methods, it only cares about the pairwise correlation between heterogeneous multi-omics endophenotypes. In biomedical studies, the high-order association among multi-omics 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 high-order endophenotype correlation guided MTSCCA (*hoc*MTSCCA), which is defined as

$$\begin{aligned}\max_{\mathrm{U},\mathrm{v}_i}\sum_{i=1}^{3}\mathrm{u}_i^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}_i\mathrm{v}_i+\Lambda\mathrm{corr}\left(\mathrm{Y}_1\mathrm{v}_1,\mathrm{Y}_2\mathrm{v}_2,\mathrm{Y}_3\mathrm{v}_3\right)\\ -R\left(\mathrm{U}\right)-\sum_i R\left(\mathrm{v}_i\right)\\ s.t.\left\|\mathrm{X}\mathrm{u}_i\right\|_2^2=1,\left\|\mathrm{Y}_i\mathrm{v}_i\right\|_2^2=1,\forall i=1,2,3.\end{aligned}\tag{17}$$

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

$$\begin{aligned}\min_{\mathrm{U},\mathrm{v}_i}\sum_{i=1}^{3}\text{-}\mathrm{u}_i^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y}_i\mathrm{v}_i-\Lambda C_{123}\,\bar{\times}_1\,\mathrm{v}_1\,\bar{\times}_2\,\mathrm{v}_2\,\bar{\times}_3\,\mathrm{v}_3\\ +R\left(\mathrm{U}\right)+\sum_i R\left(\mathrm{v}_i\right)\\ s.t.\left\|\mathrm{X}\mathrm{u}_i\right\|_2^2=1,\left\|\mathrm{Y}_i\mathrm{v}_i\right\|_2^2=1,\forall i=1,2,3.\end{aligned}\tag{18}$$

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

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

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

This high-order canonical correlation quantifies the CEP cross multi-omics 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-omics 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 association, while *pc*MTSCCA emphasizes on the pairwise CEP association. 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 modelling strategy to *pc*MTSCCA, it can be solved in the same way as presented in Algorithm 1, and its convergence is also guaranteed.

Conclusion

Alzheimer's disease (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-omics data, to comprehensively identify genetic loci. Most existing methods cannot make use of both multi-omics endophenotypes and their CEP associations, and thus are suboptimal. Therefore, we proposed two integrated MTSCCA 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-omics methods, i.e., SMCCA, Adaptive SMCCA and RelPMDCCA on both simulated and real world data. The results of synthetic 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. But comparison methods either yielded too many or too little markers which was undesirable for real biomedical studies. Our methods provided very helpful clues for further in-depth investigation and thus were powerful in multi-omics heterogeneous markers identification. In the future, we will apply both methods to large scales data sets which is important for whole genome sequence analysis.

Code availability

inMTSCCA is implemented in Matlab, and the usage and manual of inMTSCCA is also available at BioCode at National Genomics Data Center: https://ngdc.cncb.ac.cn/biocode/tools/BT007330. Non-commercial use for academic, government, and non-profit institutions is permitted.

CRediT author statement

Du L: Conceptualization, Methodology, Writing- Original draft preparation. **Zhang J:** Software, Visualization, Formal analysis. **Zhao Y:** Investigation. **Shang M:** Validation. **Guo L:** Writing - Review & Editing. **Han J:** Writing - Review & Editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. HoffmannLa Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICLtd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.;Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work was supported in part by the National Key R&D Program of China [2022ZD0213700]; National Natural Science Foundation of China [61973255, 61936007, 62136004]; Natural Science Basic Research Program of Shaanxi [2020JM-142] and China Postdoctoral Science Foundation [2020T130537] at Northwestern Polytechnical University.

ORCID

0000-0002-6698-9814 (Lei Du).

0000-0003-1688-5547 (Jin Zhang).

0000-0002-3215-9986 (Ying Zhao).

0000-0002-9227-7263 (Muheng Shang).

0000-0003-0728-896X (Lei Guo).

0000-0001-5545-7217 (Junwei Han).

References

[1] Sims R, Hill M, Williams JJNn. The multiplex model of the genetics of Alzheimer's disease. Nat Neurosci 2020;23:311-22.

[2] Andrews SJ, Fulton-Howard B, Goate AJTLN. 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 JWJNRG. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet 2013;14:483-95.

[4] Casale FP, Rakitsch B, Lippert C, Stegle OJNm. Efficient set tests for the genetic analysis of correlated traits. Nat Methods 2015;12:755-8.

[5] Guo B, Wu BJB. 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 V, 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 endophenotype 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 G, Insel P, Tosun D, Schuff N, Truran-Sacrey D, Raptentsetsang S, et al. Hippocampal atrophy rates and CSF biomarkers in elderly APOE2 normal subjects. Neurology 2010;75:1976-81.

[11] Tyler AL, Crawford DC, Pendergrass SAJBib. The detection and characterization of pleiotropy: discovery, progress, and promise. Brief Bioinform 2016;17:13-22.

[12] Shen L, Thompson PMJPotI. Brain imaging genomics: integrated analysis and machine learning. Proc IEEE Inst Electr Electron Eng 2019;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;8:1-12.

[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 Y-PJMia. 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-i71.

[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 RJJSaig, biology m. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol 2009;8.

[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 2017;65:390-9.

[20] Rodosthenous T, Shahrezaei V, Evangelou MJB. 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 2019;18:227-39.

[23] Witten DM, Tibshirani R, Hastie TJB. 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 YJItoK, Engineering D. Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans Knowl Data Eng 2015;27:3111-24.

[25] Kolda TG, Bader BWJSr. Tensor decompositions and applications. SIAM Rev Soc Ind Appl Math 2009;51:455-500.

[26] Chen X, Liu HJSiB. 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 JJNrr. A non-invasive, rapid method to genotype late-onset Alzheimer's disease-related apolipoprotein E gene polymorphisms. Neural Regen Res 2014;9:69.

[29] Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers H, et al. Proteome-based plasma biomarkers for Alzheimer's disease. Brain 2006;129:3042-50.

[30] Nilsson K, Gustafson L, Hultberg BJCb. 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. JAMA 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;7:1-6.

[33] Mulder SD, Hack CE, van der Flier WM, Scheltens P, Blankenstein MA, Veerhuis RJJoAsD. Evaluation of intrathecal serum amyloid P (SAP) and C-reactive protein (CRP) synthesis in Alzheimer's disease with the use of index values. JAD 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.

Figure legends

Figure 1 Canonical weights on synthetic data from 20 times five-fold cross-validation.

Each row is a method: (1) Ground truth; (2) SMCCA; (3) Adaptive SMCCA; (4) RelPMDCCA; (5) *pc*MTSCCA; and (6) *hoc*MTSCCA. Each column corresponds to a canonical weight. The first column u is the canonical weight of X, the second column v_1 is that of Y_1, the third one is the canonical weight of Y_2, and the last column is that for Y_3. The values in each panel are obtained from 20 times trials. Data source: four simulated data sets with different ground truths. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Figure 2 Canonical weights (mean value) of SNPs from 20 times fivefold cross-validation.

Each row is a method: (1) SMCCA; (2) Adaptive SMCCA; (3) RelPMDCCA; (4) *pc*MTSCCA; and (5) *hoc*MTSCCA. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; SNP, single nucleotide polymorphism;

Figure 3 Canonical weights (mean value) of plasma-derived proteomic markers from 20 times five-fold cross-validation.

Each row is a method: (1) SMCCA; (2) Adaptive SMCCA; (3) RelPMDCCA; (4) *pc*MTSCCA; and (5) *hoc*MTSCCA. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Figure 4 Canonical weights (mean value) of CSF-derived proteomic markers from 20 times five-fold cross-validation.

Each row is a method: (1) SMCCA; (2) Adaptive SMCCA; (3) RelPMDCCA; (4) *pc*MTSCCA; and (5) *hoc*MTSCCA. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; CSF, cerebrospinal fluid.

Figure 5 Canonical weights (mean value) of brain imaging QTs from 20 times five-fold cross-validation.

Each row is a method: (1) SMCCA; (2) Adaptive SMCCA; (3) RelPMDCCA; (4) *pc*MTSCCA; and (5) *hoc*MTSCCA. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; QT, quantitative traits.

Figure 6 *pc*MTSCCA's pairwise correlation.

A. The association between identified SNPs and plasma-derived proteomic biomarkers. **B.** The association between identified SNPs and CSF-derived proteomic biomarkers. **C.** The association between identified SNPs and imaging QTs. The '×' symbol indicated that this pairwise association reached the significance level ($p < 0.05$). pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis.

Figure 7 *hoc*MTSCCA's pairwise correlation.

A. The association between identified SNPs and plasma-derived proteomic biomarkers. **B.** The association between identified SNPs and CSF-derived proteomic biomarkers. **C.** The association between identified SNPs and imaging QTs. The '×' symbol indicated that this pairwise association reached the significance level ($p < 0.05$). hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Figure 8 Pairwise comparisons for SNP and endophenotype among different diagnostic groups such as HC, MCI and AD respectively.

A. The concentration of plasma-APOE for different genotypes among three groups.

B. The concentration of CSF-APOE for different genotypes among three groups.

C. The atrophy in the left hippocampus lobe for different genotypes among three groups. SNP, single nucleotide polymorphism; HC, healthy control; MCI, mild cognitive impairment; AD, Alzheimer's Disease.

Tables legends

Table 1 The CCCs (mean ± std.) from 20 times five-fold cross-validation on synthetic data sets. CCC, canonical correlation coefficient.

Table 2 Participant characteristics.

Table 3 Comparison of the CCCs (mean ± std.) from 20 times five-fold cross-validation on ADNI. The best values were shown in bold. CCC, canonical correlation coefficient.

Table 4 Top ten loci of each method based on mean canonical weights.

Table 5 Top ten plasma-derived proteomic markers of each method based on mean canonical weights.

Table 6 Top ten CSF-derived proteomic markers of each method based on mean canonical weights. CSF, cerebrospinal fluid.

Table 7 Top ten brain imaging QTs of each method based on mean canonical weights. QTs, quantitative traits. QT, quantitative traits.

Supplementary material

Table S1 Top ten loci of each method based on mean canonical weights

Table S2 Top ten plasma-derived proteomic markers of each method based on mean canonical weights

Table S3 Top ten CSF-derived proteomic markers of each method based on mean canonical weights. CSF, cerebrospinal fluid.

Table S4 Top ten brain imaging QTs of each method based on mean canonical weights. QT, quantitative traits.

Figure 1

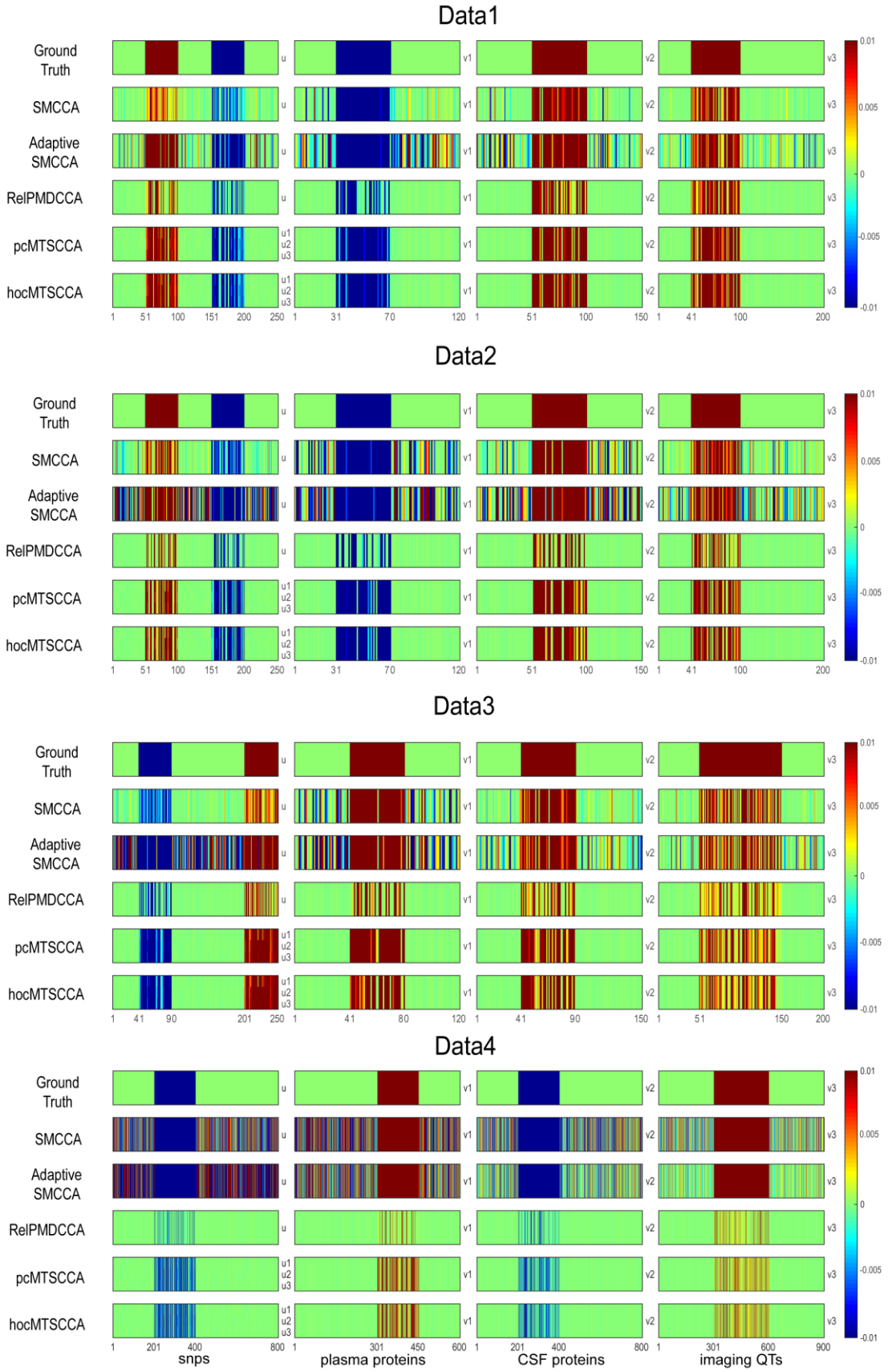
Data1
Data2
Data3
Data4
Ground Truth
SMCCA
Adaptive SMCCA
RelPMDCCA
pcMTSCCA
hocMTSCCA
u
u1
u2
u3
v1
v2
v3
0.01
0.005
0
-0.005
-0.01
snps
plasma proteins
CSF proteins
imaging QTs

Figure 2

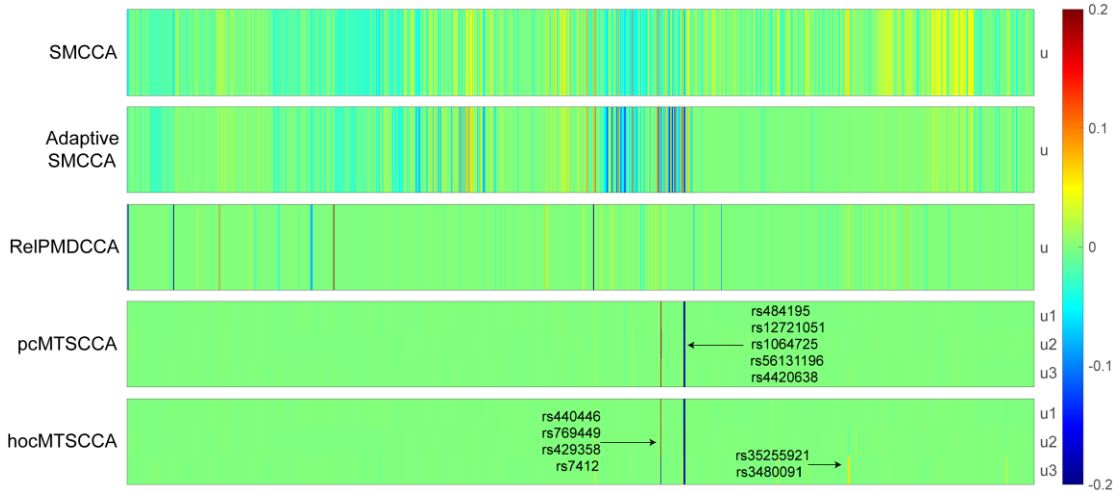

Figure 3

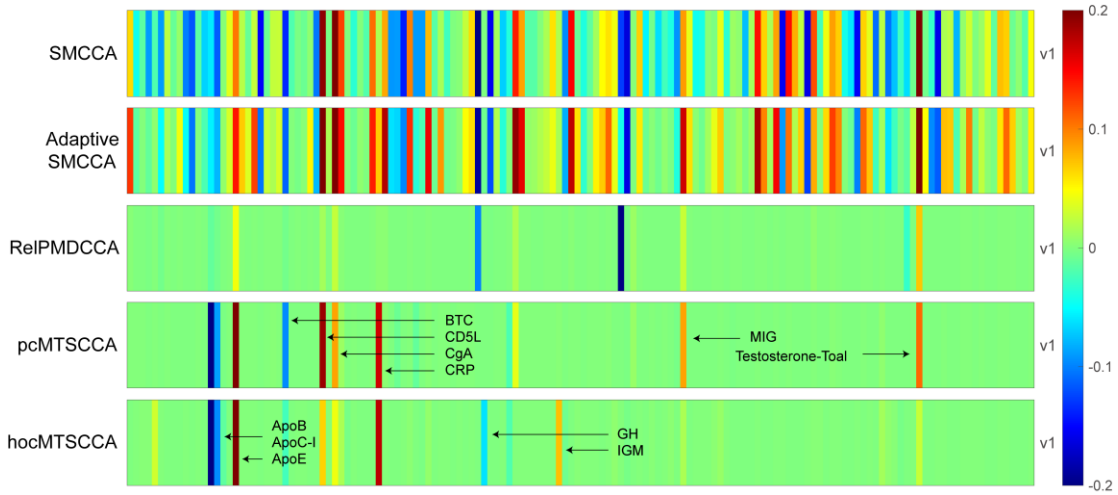

Figure 4

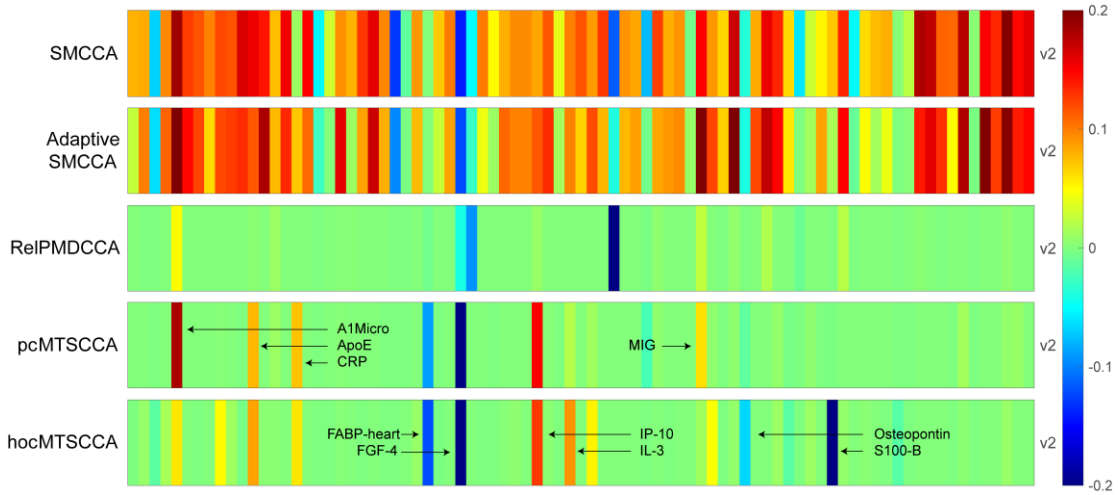

Figure 5

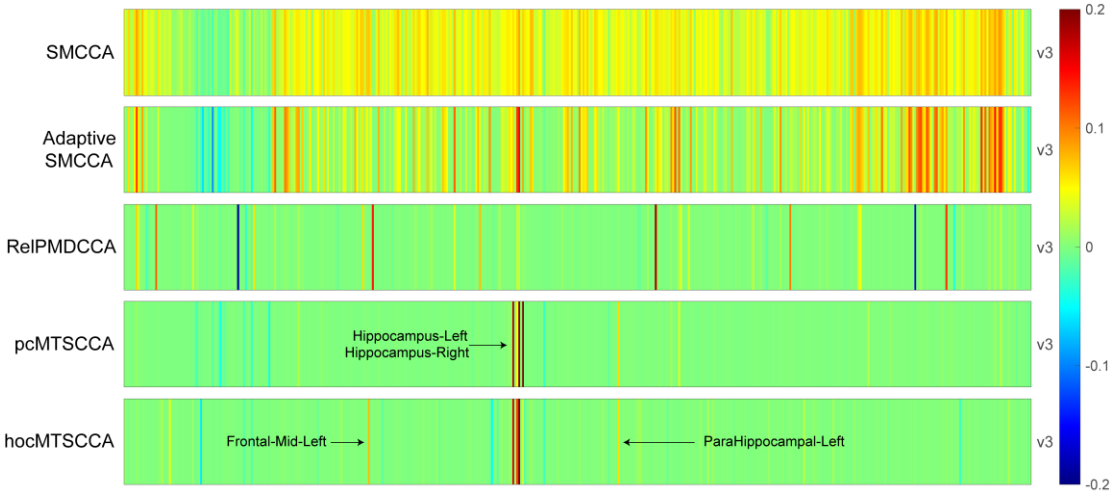

Figure 6

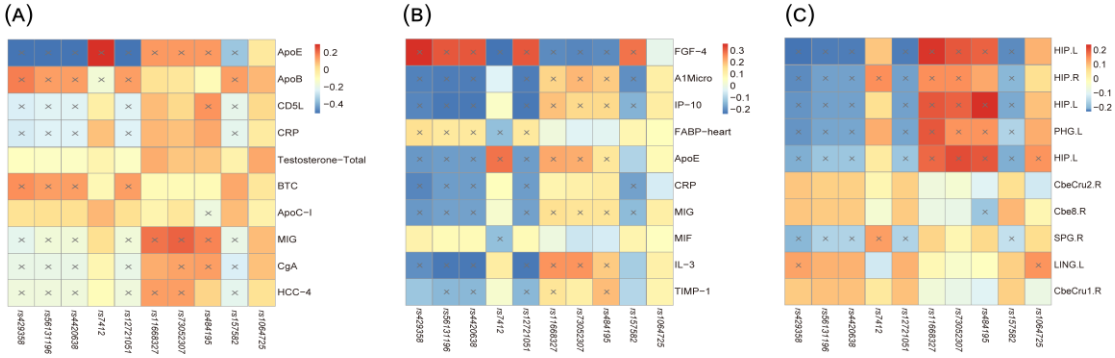

Figure 7

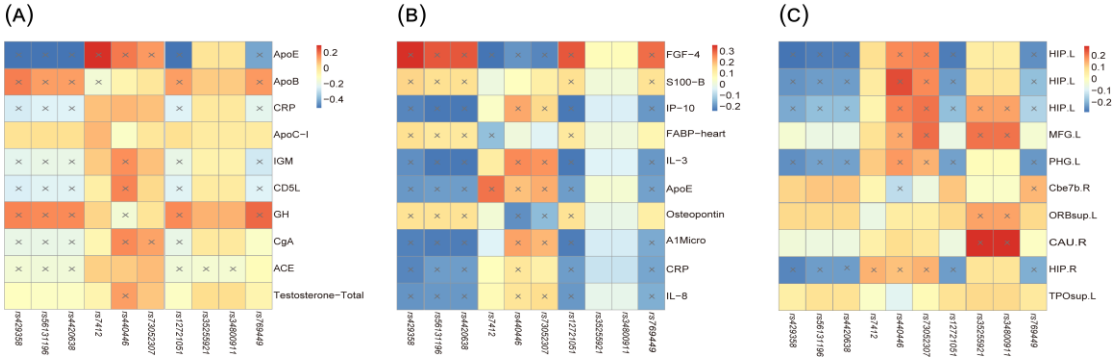

Figure 8

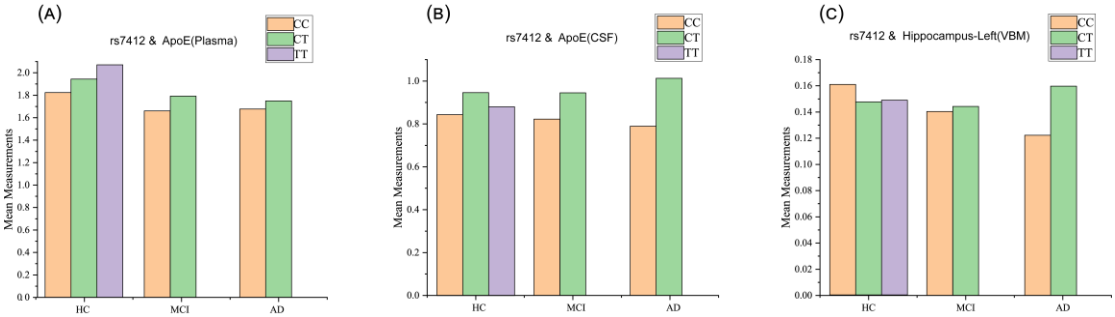

Table 1: The CCCs (mean ± std.) from 20 times five-fold cross-validation on synthetic data sets

	Data1			Data2			Data3			Data4		
	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 CCCs												
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
*Rel*PMDCCA	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 CCCs												
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
*Rel*PMDCCA	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: CCC, Canonical correlation coefficient. The CCC between X and Y1 as CCC1-1, that between X and Y2 as CCC1-2 and so on. SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Table 2: Participant characteristics

	HC	MCI	AD
Num	42	137	65
Gender (M/F, %)	52.38/47.62	69.34/30.66	55.38/44.62
Handedness (R/L, %)	90.48/9.52	92.70/7.30	98.46/1.54
Age (mean ± std)	75.40±5.80	74.13±7.22	74.75±7.67
Education (mean ± std)	15.88±2.77	16.03±2.98	15.12±3.05

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

Table 3: Comparison of the CCCs (mean ± std.) from 20 times five-fold cross-validation on ADNI. The best values were shown in bold

Training CCCs			
	SNP-Plasma	SNP-CSF	SNP-VBM
SMCCA	0.33 ± 0.04	0.33 ± 0.04	0.29 ± 0.02
Adaptive SMCCA	0.36 ± 0.03	0.34 ± 0.03	0.23 ± 0.02
*Rel*PMDCCA	0.44 ± 0.05	0.46 ± 0.03	**0.50 ± 0.04**
*pc*MTSCCA	0.66 ± 0.04	0.44 ± 0.04	0.37 ± 0.04
*hoc*MTSCCA	**0.66 ± 0.03**	**0.50 ± 0.04**	0.44 ± 0.04

Testing CCCs			
	SNP-Plasma	SNP-CSF	SNP-VBM
SMCCA	0.13 ± 0.09	0.18 ± 0.10	0.09 ± 0.06
Adaptive SMCCA	0.17 ± 0.11	0.21 ± 0.11	0.10 ± 0.07
*Rel*PMDCCA	0.13 ± 0.12	0.11 ± 0.08	0.12 ± 0.09
*pc*MTSCCA	**0.56 ± 0.11**	0.30 ± 0.12	**0.17 ± 0.11**
*hoc*MTSCCA	**0.56 ± 0.11**	**0.31 ± 0.13**	0.15 ± 0.10

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; CCC, Canonical correlation coefficient.

Table 4: Top ten loci of each method based on mean canonical weights

SMCCA	Adaptive SMCCA	*Rel*PMDCCA	*pc*MTSCCA	*hoc*MTSCCA
rs11668327	rs440446	rs11083749	rs429358	rs429358
rs440446	rs429358	rs111654618	rs56131196	rs56131196
rs73052307	rs5117	rs111740474	rs4420638	rs4420638
rs157580	rs483082	rs111766460	rs7412	rs7412
rs449647	rs438811	rs77213073	rs12721051	rs440446
rs11669338	rs12721051	rs4803776	rs11668327	rs73052307
rs11673139	rs56131196	rs41301961	rs73052307	rs12721051
rs17561351	rs4420638	rs139957871	rs484195	rs35255921
rs138235833	rs11668327	rs35106910	rs157582	rs34800911
rs41290102	rs157580	rs73045691	rs1064725	rs769449

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Table 5: Top ten plasma-derived proteomic markers of each method based on mean canonical weights

SMCCA	Adaptive SMCCA	*Rel*PMDCCA	*pc*MTSCCA	*hoc*MTSCCA
Testosterone	Testosterone	Leptin	ApoE	ApoE
CgA	CD5L	FSH	ApoB	ApoB
FSH	CgA	Testosterone	CD5L	CRP
CD5L	HCC-4	ApoE	CRP	ApoC-I
LH	FSH	TBG	Testosterone	IGM
PDGF-BB	Myoglobin	MIG	BTC	CD5L
Thrombospondin-1	Cystatin-C	CgA	ApoC-I	GH
GRO-alpha	IL-16	HCC-4	MIG	CgA
PAI-1	MIG	CD5L	CgA	ACE
RANTES	LH	ApoB	HCC-4	Testosterone

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Table 6: Top ten CSF-derived proteomic markers of each method based on mean canonical weights

SMCCA	**Adaptive SMCCA**	***Rel*PMDCCA**	***pc*MTSCCA**	***hoc*MTSCCA**
VCAM-1	A1Micro	Leptin	FGF-4	FGF-4
A1Micro	NGAL	FSH	A1Micro	S100-B
TRAIL-R3	MIG	A1Micro	IP-10	IP-10
TIMP-1	TFF3	FGF-4	FABP-heart	FABP-heart
TM	VCAM-1	MIG	ApoE	IL-3
NGAL	ApoH	SAP	CRP	ApoE
ApoD	TIMP-1	PLGF	MIG	Osteopontin
vWF	PLGF	LPa	MIF	A1Micro
C3	C3	IP-10	IL-3	CRP
PLGF	HCC-4	PRL	TIMP-1	IL-8

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; CSF, cerebrospinal fluid.

Table 7: Top ten brain imaging QTs of each method based on mean canonical weights

SMCCA	**Adaptive SMCCA**	***Rel*PMDCCA**	***pc*MTSCCA**	***hoc*MTSCCA**
STG.R	HIP.L	CbeCru1.L	HIP.L	HIP.L
ANG.R	STG.R	IPL.R	HIP.R	HIP.L
STG.R	STG.R	MTG.L	HIP.L	HIP.L
STG.R	STG.L	MFG.L	PHG.L	MFG.L
HIP.L	ANG.R	TPOmid.L	HIP.L	PHG.L
MTG.R	HIP.L	CAL.R	CbeCru2.R	Cbe7b.R
MTG.R	STG.R	PCUN.R	Cbe8.R	ORBsup.L
STG.L	MTG.L	SFGmed.L	SPG.R	CAU.R
STG.L	MTG.R	ANG.R	LING.L	HIP.R
TPOmid.L	SPG.R	MFG.L	CbeCru1.R	TPOsup.L

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; QT, quantitative traits.

Supplementary material

Table S1 Top ten loci of each method based on mean canonical weights

SMCCA		Adaptive SMCCA		RelPMDCCA		*pc*MTSCCA		*hoc*MTSCCA	
SNP_ID	Weight	SNP_ID	Weight	SNP_ID	Weight	SNP_ID	Weight	SNP_ID	Weight
rs11668327	0.10	rs440446	0.14	rs11083749	-0.31	rs429358	-0.51	rs429358	-0.29
rs440446	0.09	rs429358	-0.13	rs111654618	0.27	rs56131196	-0.19	rs56131196	-0.24
rs73052307	0.09	rs5117	-0.13	rs111740474	-0.22	rs4420638	-0.19	rs4420638	-0.23
rs157580	0.09	rs483082	-0.13	rs111766460	-0.15	rs7412	0.07	rs7412	0.04
rs449647	0.08	rs438811	-0.13	rs77213073	0.11	rs12721051	-0.02	rs440446	0.02
rs11669338	0.08	rs12721051	-0.13	rs4803776	-0.10	rs11668327	0.02	rs73052307	0.02
rs11673139	0.08	rs56131196	-0.13	rs41301961	-0.08	rs73052307	0.01	rs12721051	-0.02
rs17561351	0.07	rs4420638	-0.13	rs139957871	-0.08	rs484195	0.01	rs35255921	0.02
rs138235833	0.07	rs11668327	0.12	rs35106910	-0.08	rs157582	-0.01	rs34800911	0.02
rs41290102	0.07	rs157580	0.12	rs73045691	-0.07	rs1064725	0.01	rs769449	-0.01

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Table S2 Top ten plasma-derived proteomic markers of each method based on mean canonical weights

SMCCA		Adaptive SMCCA		RelPMDCCA		*pc*MTSCCA		*hoc*MTSCCA	
Plasma_ID	Weight	Plasma_ID	Weight	Plasma_ID	Weight	Plasma_ID	Weight	Plasma_ID	Weight
Testosterone-Total	0.24	Testosterone-Total	0.23	Leptin	-0.76	Apolipoprotein E	0.82	Apolipoprotein E	0.92
Chromogranin-A	0.21	CD5	0.23	Follicle-Stimulating Hormone	-0.10	Apolipoprotein B	-0.27	Apolipoprotein B	-0.22
Follicle-Stimulating Hormone	-0.20	Chromogranin-A	0.21	Testosterone-Total	0.07	CD5	0.19	C-Reactive Protein	0.18
CD5	0.19	Chemokine CC-4	0.19	Apolipoprotein E	0.04	C-Reactive Protein	0.17	Apolipoprotein C-I	-0.10
Luteinizing Hormone	-0.16	Follicle-Stimulating Hormone	-0.19	Thyroxine-Binding Globulin	-0.03	Testosterone-Total	0.11	Immunoglobulin M	0.07
Platelet-Derived Growth Factor BB	-0.16	Myoglobin	0.18	Monokine Induced by Gamma Interferon	0.03	Betacellulin	-0.10	CD5	0.07
Thrombospondin-1	-0.16	Cystatin-C	0.18	Chromogranin-A	0.03	Apolipoprotein C-I	-0.09	Growth Hormone	-0.06
Growth-Regulated alpha protein	-0.16	Interleukin-16	0.17	Chemokine CC-4	0.02	Monokine Induced by Gamma Interferon	0.09	Chromogranin-A	0.04
Plasminogen Activator Inhibitor 1	-0.16	Monokine Induced by Gamma Interferon	0.17	CD5	0.01	Chromogranin-A	0.08	Angiotensin-Converting Enzyme	0.03
T-Cell-Specific Protein RANTES	-0.15	Luteinizing Hormone	-0.16	Apolipoprotein B	-0.01	Chemokine CC-4	0.04	Testosterone-Total	0.03

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA.

Table S3 Top ten CSF-derived proteomic markers of each method based on mean canonical weights

SMCCA		Adaptive SMCCA		RelPMDCCA		*pc*MTSCCA		*hoc*MTSCCA	
CSF_ID	Weight	CSF_ID	Weight	CSF_ID	Weight	CSF_ID	Weight	CSF_ID	Weight
Vascular Cell Adhesion Molecule-1	0.20	Alpha-1-Microglobulin	0.26	Leptin	-0.79	Fibroblast Growth Factor 4	-0.68	Fibroblast Growth Factor 4	-0.65
Alpha-1-Microglobulin	0.19	Neutrophil Gelatinase-Associated Lipocalin	0.25	Follicle-Stimulating Hormone	-0.09	Alpha-1-Microglobulin	0.18	S100 calcium-binding protein B	-0.21
TRAIL receptor 3	0.18	Monokine Induced by Gamma Interferon	0.20	Alpha-1-Microglobulin	0.05	Interferon gamma Induced Protein 10	0.15	Interferon gamma Induced Protein 10	0.13
Tissue Inhibitor of Metalloproteinases 1	0.18	Trefoil Factor 3	0.20	Fibroblast Growth Factor 4	-0.04	Fatty Acid-Binding Protein-heart	-0.09	Fatty Acid-Binding Protein-heart	-0.12
Thrombomodulin	0.17	Vascular Cell Adhesion Molecule-1	0.20	Monokine Induced by Gamma Interferon	0.02	Apolipoprotein E	0.08	Interleukin-3	0.09
Neutrophil Gelatinase-Associated Lipocalin	0.17	Apolipoprotein H	0.18	Serum Amyloid P-Component	0.02	C-Reactive Protein	0.07	Apolipoprotein E	0.09
.09Apolipoprotein D	0.17	Tissue Inhibitor of Metalloproteinases 1	0.18	Placenta Growth Factor	0.02	Monokine Induced by Gamma Interferon	0.06	Osteopontin	-0.07
von Willebrand Factor	0.16	Placenta Growth Factor	0.17	Lipoprotein(a)	0.01	Macrophage Migration Inhibitory Factor	-0.02	Alpha-1-Microglobulin	0.06
Complement C3	0.16	Complement C3	0.17	Interferon gamma Induced Protein 10	0.01	Interleukin-3	0.02	C-Reactive Protein	0.06
Placenta Growth Factor	0.16	Chemokine CC-4	0.16	Prolactin	-0.01	Tissue Inhibitor of Metalloproteinases 1	0.01	Interleukin-8	0.05

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; CSF, cerebrospinal fluid.

Table S4 Top ten brain imaging QTs of each method based on mean canonical weights

SMCCA		Adaptive SMCCA		RelPMDCCA		*pc*MTSCCA		*hoc*MTSCCA	
ROI_ID	Weight	ROI_ID	Weight	ROI_ID	Weight	ROI_ID	Weight	ROI_ID	Weight
Temporal Sup Right	0.09	Hippocampus Left	0.15	Cerebelum Crus1 Left	-0.19	Hippocampus Left	0.45	Hippocampus Left	0.36
Angular Right	0.09	Temporal Sup Right	0.15	Parietal Inf Right	0.18	Hippocampus Right	0.23	Hippocampus Left	0.16
Temporal Sup Right	0.09	Temporal Sup Right	0.13	Temporal Mid Left	-0.17	Hippocampus Left	0.17	Hippocampus Left	0.13
Temporal Sup Right	0.08	Temporal Sup Left	0.13	Frontal Mid Left	0.14	ParaHippocampal Left	0.06	Frontal Mid Left	0.08
Hippocampus Left	0.08	Angular Right	0.13	Temporal Pole Mid Left	0.13	Hippocampus Left	0.05	ParaHippocampal Left	0.06
Temporal Mid Right	0.08	Hippocampus Left	0.13	Calcarine Right	0.11	Cerebelum Crus2 Right	-0.05	Cerebelum 7b Right	-0.06
Temporal Mid Right	0.08	Temporal Sup Right	0.13	Precuneus Right	0.10	Cerebelum 8 Right	-0.05	Frontal Sup Orb Left	-0.05
Temporal Sup Left	0.08	Temporal Mid Left	0.12	Frontal Sup Medial Left	0.07	Parietal Sup Right	0.03	Caudate Right	0.04
Temporal Sup Left	0.08	Temporal Mid Right	0.12	Angular Right	0.06	Lingual Left	-0.03	Hippocampus Right	0.04
Temporal Pole Mid Left	0.08	Parietal Sup Right	0.12	Frontal Mid Left	0.06	Cerebelum Crus1 Right	-0.03	Temporal Pole Sup Left	-0.03

Note: SMCCA, Sparse multiple canonical correlation analysis; RelPMDCCA, Relaxed penalised matrix decomposition CCA; pcMTSCCA, pairwise endophenotype correlation guided multi-task sparse canonical correlation analysis; hocMTSCCA, high-order endophenotype correlation guided MTSCCA; QT, quantitative traits.