Computational and Structural Biotechnology Journal 23 (2024) 3288–3299

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Identification of genetic basis of brain imaging by group sparse multi-task learning leveraging summary statistics

Duo Xi, Dingnan Cui, Mingjianan Zhang, Jin Zhang, Muheng Shang, Lei Guo, Junwei Han *, Lei Du *

Northwestern Polytechnical University, Xi'an, 710072, China

ARTICLE INFO

Keywords:
Brain imaging genetics
Machine learning
Sparse multi-task learning
Feature selection
Summary statistics

ABSTRACT

Brain imaging genetics is an evolving neuroscience topic aiming to identify genetic variations related to neuroimaging measurements of interest. Traditional linear regression methods have shown success, but their reliance on individual-level imaging and genetic data limits their applicability. Herein, we proposed S-GsMTLR, a group sparse multi-task linear regression method designed to harness summary statistics from genome-wide association studies (GWAS) of neuroimaging quantitative traits. S-GsMTLR directly employs GWAS summary statistics, bypassing the requirement for raw imaging genetic data, and applies multivariate multi-task sparse learning to these univariate GWAS results. It amalgamates the strengths of conventional sparse learning methods, including sophisticated modeling techniques and efficient feature selection. Additionally, we implemented a rapid optimization strategy to alleviate computational burdens by identifying genetic variants associated with phenotypes of interest across the entire chromosome. We first evaluated S-GsMTLR using summary statistics derived from the Alzheimer's Disease Neuroimaging Initiative. The results were remarkably encouraging, demonstrating its comparability to conventional methods in modeling and identification of risk loci. Furthermore, our method was evaluated with two additional GWAS summary statistics datasets: One focused on white matter microstructures and the other on whole brain imaging phenotypes, where the original individual-level data was unavailable. The results not only highlighted S-GsMTLR's ability to pinpoint significant loci but also revealed intriguing structures within genetic variations and loci that went unnoticed by GWAS. These findings suggest that S-GsMTLR is a promising multivariate sparse learning method in brain imaging genetics. It eliminates the need for original individual-level imaging and genetic data while demonstrating commendable modeling and feature selection capabilities.

1. Introduction

Brain imaging genetics is an influential and captivating brain science field. Generally, it jointly analyzes genetic variations (e.g., single nucleotide polymorphisms, SNPs), structural or functional neuroimaging scans (e.g., quantitative traits, QTs) [1,2]. A primary goal of imaging genetics is to investigate the associations between brain imaging QTs and SNPs, with the expectation of uncovering the genetic basis of brain structures and disorders [3,4].

Multi-task linear regression (MTLR) is a widely adopted sparse learning method in brain imaging genetics. Unlike univariate linear regression, MTLR simultaneously examines the effects of genetic variants on multiple neuroimaging quantitative traits (QTs) [5–8]. By exploring various types of sparsity, MTLR facilitates the identification of significant genetic variants at different levels, such as groups of SNPs within the same gene [9–11]. Despite the success of MTLR methods, they face a significant practical challenge: conventional MTLR relies on access to original individual-level imaging and genetic data, rendering them ineffective when such data is unavailable.

Genome-wide association studies (GWAS) aims to identify genetic variants that exhibit significant associations with phenotypic traits, emerging as a widely-used method over the past decade [12,13]. Fortunately, GWAS publicly release their summary statistics results, offering a wealth of intermediate data on the associations between imaging QTs and SNPs. To date, GWAS has been widely applied to brain imaging phenotypes, successfully identifying numerous significant genetic variants associated with brain structure and disorders [14–16]. For example, GWAS has been applied to investigate the heritability of human brain

* Corresponding author.
E-mail addresses: jhan@nwpu.edu.cn (J. Han), dulei@nwpu.edu.cn (L. Du).

https://doi.org/10.1016/j.csbj.2024.08.027
Received 30 June 2024; Received in revised form 29 August 2024; Accepted 29 August 2024
Available online 3 September 2024

white matter microstructure [17], and other brain imaging-derived phenotypes (IDPs), such as volume based morphology [18]. Although GWAS has successfully identified significant genetic variants, it is primarily a univariate learning method, which may overlook complex yet meaningful associations between SNPs [19,4]. This limitation also applies to brain imaging QTs due to the gene pleiotropy, as GWAS can only analyze the associations between each image QT and genetic variation. Multi-task learning offers a promising solution to these challenges. Additionally, information on correlations between SNPs, such as linkage disequilibria (LD), is readily accessible through public databases like the 1000 Genomes Project Consortium [20], further enabling the feasibility of multi-task and "multi-SNP" analysis of GWAS data. For these reasons, increased efforts have been made to utilize these freely available summary statistics from GWAS to study the associations between multiple brain imaging and genetics. For example, metaCCA has been developed to identify relevant imaging and genetic biomarkers but not the precise genetic basic of QTs of interest [21]. Further, MTAG was proposed only aims to the traits whose GWAS estimates were correlated [22], CONFIT needs a prior which is hard to obtain in practical [23], and MTAR only considers variations of one gene each time [24,25]. Therefore, it is of significant interest and importance to develop sparse multi-task learning methods that can perform multi-task multivariate analysis of GWAS results without requiring access to the original imaging genetic data.

In this paper, we proposed a novel group-sparse multivariate multi-task linear regression based on GWAS summary statistics (S-GsMTLR) for brain imaging genetics. S-GsMTLR eliminates the need for original individual-level imaging genetic data by leveraging summary statistics from GWAS. To evaluate the advantages of S-GsMTLR, we conducted two investigations. First, we generated a dataset containing all SNPs on chromosome 19 from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The results indicated that S-GsMTLR performs comparably to conventional GsMTLR in terms of feature selection, root mean square errors (RMSE), and correlation coefficients, demonstrating its equivalent performance even without access to individual-level data. Second, we applied S-GsMTLR to two GWAS summary statistics datasets where the original individual-level data were unavailable. One focused on white matter microstructures (referred to as WMM GWAS), and the other studied whole-brain imaging phenotypes. The results showed that S-GsMTLR not only identified relevant loci found by GWAS but also uncovered interesting structural associations within SNPs that were missed by GWAS. This highlights the powerful capability of S-GsMTLR as a tool for imaging genetics, providing significant insights for multivariate multi-task analysis of univariate GWAS results, while offering substantial advantages over existing imaging genetic studies. Overall, S-GsMTLR represents a promising approach for brain imaging genetics, enabling effective analysis and discovery of genetic associations by leveraging summary statistics from large GWAS.

2. Methods

In this article, we used italic letters denote scalars, boldface lowercase letters represent column vectors, and boldface capitals for matrices. For $\mathbf{X} = (x_{ij})$, $\mathbf{x}^i$ denotes its i-th row, $\mathbf{x}_j$ for its j-th column, and $\mathbf{X}_i$ denotes the i-th matrix. $\|\mathbf{X}\|_2$ denotes the Euclidean norm, $\|\mathbf{X}\|_{2,1}$ denotes the sum of the Euclidean norms of the rows of $\mathbf{X}$, and $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2}$ denotes the Frobenius norm.

2.1. Backgrounds

Generally, Genome wide association studies (GWAS) studies the association between one brain imaging QT $\mathbf{y}_p \in \mathbb{R}^{n\times 1}$ and one SNP $\mathbf{x}_g \in \mathbb{R}^{n\times 1}$ by a linear regression model, i.e.,

$$\mathbf{y}_p = \alpha_{gp} + \mathbf{x}_g \beta_{gp} + \boldsymbol{\varepsilon}, \tag{1}$$

where α_{gp} is the y-intercept accommodating the effects of covariates (e.g., age, gender etc.). $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ represents the Gaussian noise, β_{gp} is the regression coefficient (effect size) of SNP g on QT p. As previously mentioned, summary statistics from large GWAS are typically freely available. This accessibility provides a unique opportunity to perform multivariate multi-task analysis using these summary statistics.

Sparse multi-task linear regression method is a most popular multivariate method in brain imaging genetics. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ represent the genotype matrix with n participants and d SNPs, and $\mathbf{Y} \in \mathbb{R}^{n\times c}$ denote the phenotype matrix with the same n subjects, and c is the number of traits (tasks). Then a general sparse multi-task regression model can be defined as:

$$\min_{\mathbf{W}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2 + \gamma_1 R_1(\mathbf{W}) + \gamma_2 R_2(\mathbf{W}). \tag{2}$$

$\mathbf{W} \in \mathbb{R}^{d\times c}$ is the regression coefficient matrix with each row loading the weights of all SNPs for an imaging QT. γ_1 and γ_2 are hyperparameters to balance the loss function and the regularization terms i.e., $R_1(\mathbf{W})$ and $R_2(\mathbf{W})$. Although conventional sparse multi-task regression models have been successfully applied to identifying multivariate associations in brain imaging genetics, their practical application is limited by the difficulty of acquiring individual-level data. Therefore, it is crucial to develop multivariate multi-task learning methods that can leverage freely available summary statistics from large GWAS. This approach facilitates multivariate multi-task analysis of univariate GWAS results while overcoming the limitations posed by the need for individual-level data.

2.2. Summary statistics based group-sparse multi-task regression method

To leverage summary statistics from GWAS, we first take the derivative of Eq. (2) and set it to zero according to conventional optimization techniques. Hence, the critical step to solve the conventional multi-task regression model is:

$$\begin{aligned}\mathbf{W} &= \left(\mathbf{X}^T\mathbf{X} + \gamma_1\mathbf{D}_1 + \gamma_2\mathbf{D}_2\right)^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \left(\Sigma_{XX} + \gamma_1\hat{\mathbf{D}}_1 + \gamma_2\hat{\mathbf{D}}_2\right)^{-1}\Sigma_{XY},\end{aligned} \tag{3}$$

where $\hat{\mathbf{D}}_1 = \frac{\mathbf{D}_1}{n-1}$ and $\hat{\mathbf{D}}_2 = \frac{\mathbf{D}_2}{n-1}$ are block diagonal matrices, and both of them are deduced from the gradient or sub-gradient of $R_1(\mathbf{W})$ and $R_2(\mathbf{W})$. In practice, different regularization terms could yield different levels of sparsity, leading to different subsets of SNPs of relevance. In order to identify the most meaningful genetic risk factors, we applied $G_{2,1}$-norm and l_1-norm. Specifically, $R_1(\mathbf{W}) = \|\mathbf{W}\|_{G_{2,1}} = \sum_{k=1}^{K}\sqrt{\sum_{i\in n_k}\sum_{j=1}^{c} w_{ij}^2}$, where SNPs are grouped into K groups by gene, π_k for $k = 1, \ldots, K$ represents the k-th set of SNPs. Then the kth diagonal block of $\hat{\mathbf{D}}_1$ is $\frac{1}{2(n-1)\|\mathbf{W}^k\|_F}$ where $\mathbf{I}_k$ is an identity matrix whose size is π_k, and the i-th diagonal element of $\hat{\mathbf{D}}_2$ is $\frac{1}{2(n-1)\|\mathbf{W}^i\|_2}$.

Moreover, the two most important components of the formula are $\Sigma_{XX} = \frac{\mathbf{X}^T\mathbf{X}}{n-1}$ and $\Sigma_{XY} = \frac{\mathbf{X}^T\mathbf{Y}}{n-1}$, which represent the intra-covariance of genotype data and the inter-covariance between genotype and phenotypic data, respectively. Consequently, once Σ_{XY} and Σ_{XX} are obtained, the solution $\mathbf{W}$ can be determined. Therefore, we will present the solutions for Σ_{XY} and Σ_{XX} in the following paragraphs without requiring access the original genetic matrix $\mathbf{X}$ and imaging data $\mathbf{Y}$.

2.2.1. Calculating the inter-covariance Σ_{XY}

According to Eq. (1), when we normalize the genetype and phenotype matrices with zero mean and unit variance respectively, the effect size β_{gp} would equal to their covariance, i.e.:

$$\beta_{gp} = \left(\mathbf{x}_g^T\mathbf{y}_p\right)\left(\mathbf{x}_g^T\mathbf{x}_g\right)^{-1} = \frac{\mathbf{x}_g^T\mathbf{y}_p}{n-1}. \tag{4}$$

Therefore, we can calculate the inter-covariance $\Sigma_{XY} \in \mathbb{R}^{d\times c}$ in Eq. (3) with a corresponding summary statistics matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times c}$, i.e.,

$$\Sigma_{XY} = \frac{\mathbf{X}^T\mathbf{Y}}{n-1} = \begin{pmatrix} \frac{\mathbf{x}_1\mathbf{y}_1}{n-1} & \cdots & \frac{\mathbf{x}_1\mathbf{y}_c}{n-1} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{x}_d\mathbf{y}_1}{n-1} & \cdots & \frac{\mathbf{x}_d\mathbf{y}_c}{n-1} \end{pmatrix} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1c} \\ \vdots & \ddots & \vdots \\ \beta_{d1} & \cdots & \beta_{dc} \end{pmatrix}. \tag{5}$$

In practice, we do not know that whether the summary statistics are normalized, thus we suggest transforming β_{gp} first, i.e., $\beta_{gp}^{\text{normalized}} = \frac{1}{\sqrt{N}se_{gp}} \times \beta_{gp}$, where se_{gp} represents the standard error of β_{gp}, and N denotes the sample size [21]. Consequently, we can leverage the widely published summary statistics from large GWAS to obtain the inter-covariance Σ_{XY} in Eq. (3) easily.

2.2.2. Calculating the intra-covariance Σ_{XX}

We now need to find a way to calculate the intra-covariance Σ_{XX}. This intra-covariance measures the pairwise relationship of loci, indicating the genetic structure of the population. Thank to the genome stability, the intra-covariance for two subgroups from the same ethnicity could be close enough as long as the number of subjects for each subgroup is large enough [26]. Given this proposition, although we do not have the original genotype data, we can obtain a pretty good alternative if a large sub-population from the same ethnic group is available, i.e., we can effectively estimate Σ_{XX} as

$$\hat{\Sigma}_{XX} = \frac{\mathbf{X}_{ref}^T\mathbf{X}_{ref}}{n_{ref}-1}, \tag{6}$$

where $\mathbf{X}_{ref} \in \mathbb{R}^{n_{ref}\times d}$ denote the genetype data of the reference coherent with n_{ref} individuals.

The 1000 Genome Project (1kGP) database (www.internationalgenome.org) publicly provides many ethnic populations, which is usually used as reference panel. In this study, we carefully choose subjects from the same ethnic population, and calculate an approximate intra-covariance. That is, we can obtain Σ_{XX} without accessing the original genotype data $\mathbf{X}$. This approximate intra-covariance works quite well empirically and experimentally in practice.

Finally, all the building blocks of Eq. (3) have been addressed, enabling us to solve the conventional group-sparse multi-task linear regression model using only summary statistics from GWAS, which we term S-GsMTLR. For clarity, we refer to the corresponding conventional regression model we named as GsMTLR. Specifically, the iterative optimization procedure for S-GsMTLR is detailed in Algorithm 1.

Algorithm 1 The S-GsMTLR algorithm.

Require: Summary statistics matrix from GWAS, hyperparameters γ_1 and γ_2;
Output: $\hat{\mathbf{W}} \in \mathbb{R}^{d\times c}$.
1: Construct the inter-covariance matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times c}$ according to Eq. (5);
2: Calculate the intra-covariance matrix $\hat{\Sigma}_{XX} \in \mathbb{R}^{d\times d}$ from the reference haplotype data according to Eq. (6);
3: Initialize $\hat{\mathbf{W}}_1 \in \mathbb{R}^{d\times c}$ and let $t = 1$;
4: **while** not converge **do**
5: Update $\hat{\mathbf{W}}_t$ by $\hat{\mathbf{W}}_t = \left(\hat{\Sigma}_{XX} + \gamma_1\hat{\mathbf{D}}_1 + \gamma_2\hat{\mathbf{D}}_2\right)^{-1}\boldsymbol{\beta}$;
6: $t = t + 1$;
7: **end while**

2.3. Extension to chromosome-wide analysis

When applying S-GSMTLR for whole-chromosome analysis, the computation of $\hat{\Sigma}_{XX}$ becomes computationally intensive and unfeasible. However, estimating Σ_{XX} is a crucial step in solving S-GSMTLR (Algorithm 1). To address this challenge, we introduce a heuristic decoupled computational approach that employs a divide-and-conquer strategy within chromosomes to reduce the computational burden. Specifically, we divide the high-dimensional genotype matrix containing all genetic variants on the entire chromosome into smaller independent submatrices, each corresponding to the size of a LD block. Thus, the high-dimensional SNPs dataset is divided into M disjoint subsets: $\mathbf{X} = \bigoplus_{m=1}^{M}\mathbf{X}^m$, where $\bigoplus$ represents the matrix operator. This strategy leverages the inherent block structure of SNP datasets to maintain the performance of our model while decoupling and processing the interaction terms between SNPs in parallel. Consequently, we can derive the following closed solution for our proposed S-GSMTLR:

$$\hat{\mathbf{W}}_t^m = \left(\hat{\Sigma}_{XX}^m + \gamma_1\hat{\mathbf{D}}_1^m + \gamma_2\hat{\mathbf{D}}_2^m\right)^{-1}\boldsymbol{\beta},\ for\ m = 1, \ldots, M, \tag{7}$$

where $\hat{\Sigma}_{XX}^m$, $\hat{\mathbf{D}}_1^m$ and $\hat{\mathbf{D}}_2^m$ are the m-th corresponding sub-matrix of $\hat{\Sigma}_{XX}$, $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$ respectively. Similarly, this strategy can also be applied to conventional GsMTLR for chromosome-wide analysis.

2.4. Theoretical guarantees

We have the following theorem to guarantee an appropriate solution.

Theorem 1. *The difference between the solutions of S-GsMTLR and conventional GsMTLR is upper bounded by*

$$\frac{\left\|\hat{\mathbf{W}} - \mathbf{W}\right\|}{\left\|\hat{\mathbf{W}}\right\|} \le \frac{\left\|\hat{\Sigma}_{XX} - \Sigma_{XX}\right\|}{\left\|\Sigma_{XX}\right\|}, \tag{8}$$

where $\hat{\mathbf{W}}$ is the solution of S-GsMTLR and $\mathbf{W}$ is that of conventional GsMTLR.

Without loss of generality, we consider that there is only one regularization term, e.g., ℓ_1-norm, in the convectional GsMTLR model. Based on the solutions of S-GsMTLR (Algorithm 1) and conventional method (Eq. (3)), we have

$$\hat{\mathbf{W}} = \left(\hat{\Sigma}_{XX} + \gamma_1\hat{\mathbf{D}}_1\right)^{-1}\Sigma_{XY} \ \Rightarrow\ \hat{\Sigma}_{XX}\hat{\mathbf{W}} + \gamma_1\hat{\mathbf{D}}_1\hat{\mathbf{W}} = \Sigma_{XY}, \tag{9}$$

$$\mathbf{W} = \left(\Sigma_{XX} + \gamma_2\hat{\mathbf{D}}_1\right)^{-1}\Sigma_{XY} \ \Rightarrow\ \Sigma_{XX}\mathbf{W} + \gamma_2\hat{\mathbf{D}}_2\mathbf{W} = \Sigma_{XY}, \tag{10}$$

where $\hat{\Sigma}_{XX}$ and Σ_{XX} are the within-covariance obtained from the reference genotype data and the individual-level data respectively, $\hat{\mathbf{W}}$ and $\mathbf{W}$ are the regression coefficient of our S-GsMTLR and conventional GsMTLR respectively, γ_1 and γ_2 are tuning parameters of S-GsMTLR and conventional GsMTLR respectively, $\hat{\mathbf{D}}_1$ is a block diagonal matrix where the i-th diagonal block is $\frac{1}{(n-1)|\mathbf{W}^j|}(j \in [1, d])$, and $\hat{\mathbf{D}}_2$ is a block diagonal matrix where the j-th diagonal block is $\frac{1}{(n-1)|\hat{\mathbf{W}}^i|}(i \in [1, d])$. Since the final solution of $\mathbf{W}$ and $\hat{\mathbf{W}}$ are the weights of genetic data, its positive or negative values would not affect the identification results of our or conventional learning model. Therefore, we assume that the weights $\mathbf{W}$ and $\hat{\mathbf{W}}$ of different SNPs are all positive (or negative). Hence, we can get that $\hat{\mathbf{D}}_1\hat{\mathbf{W}} = \hat{\mathbf{D}}_2\mathbf{W}$ equals to the $d \times c$ matrix of $1/(n-1)$ (or $-1/(n-1)$), Eq. (9) and Eq. (10) are respectively can be turned into

$$\left(\Sigma_{XX} + \hat{\Sigma}_{XX} - \Sigma_{XX}\right)\hat{\mathbf{W}} + \frac{\gamma_1}{n-1} = \Sigma_{XY}, \tag{11}$$

$$\Sigma_{XX}\mathbf{W} + \frac{\gamma_2}{n-1} = \Sigma_{XY}. \tag{12}$$

When we subtract Eq. (12) from Eq. (11), and arrive at

$$\Sigma_{XX}\left(\hat{\mathbf{W}} - \mathbf{W}\right) + \left(\hat{\Sigma}_{XX} - \Sigma_{XX}\right)\hat{\mathbf{W}} + \frac{\gamma_1 - \gamma_2}{n-1} = 0. \tag{13}$$

Since γ_1 and γ_2 are hyperparameters and they should be the same for both models. Hence, we finally have

$$\begin{aligned} &\Sigma_{XX}\left(\hat{\mathbf{W}} - \mathbf{W}\right) = -\left(\hat{\Sigma}_{XX} - \Sigma_{XX}\right)\hat{\mathbf{W}} \\ \Rightarrow\ &\frac{\left\|\hat{\mathbf{W}} - \mathbf{W}\right\|}{\left\|\hat{\mathbf{W}}\right\|} = -\frac{\left\|\hat{\Sigma}_{XX} - \Sigma_{XX}\right\|}{\left\|\Sigma_{XX}\right\|} \end{aligned} \tag{14}$$

$$\Rightarrow \frac{\left\|\hat{\mathbf{W}}-\mathbf{W}\right\|}{\left\|\hat{\mathbf{W}}\right\|} \leq \frac{\left\|\hat{\Sigma}_{XX}-\Sigma_{XX}\right\|}{\left\|\Sigma_{XX}\right\|}.$$

Equation (14) tells us that the difference between regression coefficient of S-GsMTLR $\hat{\mathbf{W}}$ and conventional GsMTLR $\mathbf{W}$ mainly depends on the difference between estimated covariance $\hat{\Sigma}_{XX}$ and original Σ_{XX}. In other words, the closer the estimated $\hat{\Sigma}_{XX}$ being to the original Σ_{XX}, the more accurate the S-GsMTLR result is. Therefore, we suggest choosing the reference panel from the same or at least near population [26]. This can ensure a good estimation in practice. Specifically, in this paper, we estimated $\hat{\Sigma}_{XX}$ by the European population reference (EUR) haplotype data from the 1kGP as we are interested in non-Hispanic Caucasian subjects (see section 3).

3. Experiments and results

In this section, we conducted two experiments to evaluate the performance of the proposed S-GsMTLR. First, to compare the performance of our proposed model with the conventional GsMTLR, we used a dataset from the ADNI database, which includes individual-level imaging and genetic data. Specifically, we directly applied the conventional GsMTLR to the ADNI dataset. For S-GsMTLR, we first performed a univariate GWAS on the same ADNI dataset using PLINK v1.9 [27] to obtain GWAS summary statistics and then applied S-GsMTLR to these GWAS results.

In the second experiment, to assess whether our approach can identify meaningful biomarkers when only GWAS results are available, we applied S-GsMTLR to two different GWAS-only datasets, where the original individual-level data was unavailable. Since the conventional GsMTLR cannot handle GWAS-only dataset, we compare our results with previous GWAS findings in the second experiment. The results were then compared with the GWAS findings.

3.1. Experimental settings

For tuning parameters, we conducted a two-step grid search strategy to fine tune the parameters. We first searched three candidates from a moderate interval $10^i (i = [-3, -2, ..., 2, 3])$, since too large and too small parameters will lead to undesirable feature subsets. Once the suitable parameters Γ were obtained in this interval, we then search in a much smaller interval $\Gamma \pm [0.5, 0.6, ..., 5, 6]$. In the context of S-GsMTLR, due to the unavailability of individual-level data, we partitioned the summary statistics into two distinct sub-datasets for training and testing, thereby bypassing the need for individual-level data [28–30]. In the end, the stopping criterion for S-GsMTLR is $\|\mathbf{W}_{t+1}-\mathbf{W}_t\|_F / \max\left(\|\mathbf{W}_t\|_F, 1\right) \leq 10^{-4}$ in our experiments.

The ADNI data were from non-Hispanic Caucasian participants, and the subjects of IDP GWAS and WMM GWAS were European ancestry (as detailed in section 3.3). Consequently, we employed the genetic data of all 503 European individuals from 1kGP (release No. 20130521) as the reference cohort to estimate the genotypic correlation covariance Σ_{XX} in our study. The 1kGP database comprehensively collected genetic variations by sequencing the whole genome for different ethnic populations, providing a valuable public genomic resource [31].

Additionally, all experiments carried out on the same software platform and used the same data partition of the same database to ensure the fairness of comparison.

3.2. Evaluative criteria

We took widely used RMSE as the criteria to evaluate conventional GsMTLR and S-GsMTLR. Of note, the individual-level data was invisible to S-GsMTLR, and thus the conventional RMSE cannot be computed as usual. In this paper, we derived a proximate evaluation criterion instead, which can evaluate our proposed method as well.

Table 1
Participant characteristics.

	HC	MCI	AD
Number	182	292	281
Gender (M/F, %)	48.90/51.10	48.63/51.37	53.38/46.62
Handedness (R/L, %)	89.56/10.44	88.70/11.30	90.39/9.61
Age (mean±std)	73.93±5.51	70.90±6.84	72.61±8.15
Education (mean±std)	16.43±2.68	16.18±2.68	15.95±2.82

In general, for the j-th imaging QT, we calculate its RMSE value based on the ground truth $\mathbf{y}_j \in \mathbb{R}^{n\times 1}$ and its predicted value $\hat{\mathbf{y}}_j \in \mathbb{R}^{n\times 1}$ by

$$\begin{aligned}\mathrm{RMSE}\left(\mathbf{y}_j, \hat{\mathbf{y}}_j\right) &= \sqrt{\left(\|\mathbf{y}_j-\hat{\mathbf{y}}_j\|_2^2\right)/n} \\ &= \sqrt{\left(\mathbf{y}_j^T\mathbf{y}_j + \hat{\mathbf{y}}_j^T\hat{\mathbf{y}}_j - 2\mathbf{y}_j^T\hat{\mathbf{y}}_j\right)/n}.\end{aligned} \tag{15}$$

Since $\hat{\mathbf{y}}_j = \mathbf{X}\mathbf{w}_j$, we have

$$\begin{aligned}\mathrm{RMSE}\left(\mathbf{y}_j, \hat{\mathbf{y}}_j\right) &= \sqrt{\left(\mathbf{y}_j^T\mathbf{y}_j + \mathbf{w}_j\mathbf{X}^T\mathbf{X}\mathbf{w}_j - 2\left(\mathbf{X}^T\mathbf{y}_j\right)^T\mathbf{w}_j\right)/n} \\ &= \sqrt{\left(\mathbf{y}_j^T\mathbf{y}_j + \mathbf{w}_j^T\mathbf{X}^T\mathbf{X}\mathbf{w}_j - 2\left(\mathbf{X}^T\mathbf{Y}\right)_j^T\mathbf{w}_j\right)/n} \\ &= \sqrt{\left(\frac{\mathbf{y}_j^T\mathbf{y}_j}{n-1} + \mathbf{w}_j^T\Sigma_{XX}\mathbf{w}_j - 2\Sigma_{XY}{}_j^T\mathbf{w}_j\right)/\left(\frac{n}{n-1}\right)},\end{aligned} \tag{16}$$

where $\mathbf{w}_j \in \mathbb{R}^{d\times 1}$ is the weights for imaging QT j. Since the phenotypic vector $\mathbf{y}_j$ had been normalized with zero mean and unit variance, we know $\frac{\mathbf{y}_j^T\mathbf{y}_j}{n-1} = 1$. Hence, we finally had the RMSE for conventional GsMTLR as

$$\mathrm{RMSE}\left(\mathbf{y}_j, \hat{\mathbf{y}}_j\right) = \sqrt{(1-\tfrac{1}{n})\left(1 + \mathbf{w}_j^T\Sigma_{XX}\mathbf{w}_j - 2(\Sigma_{XY})_j^T\mathbf{w}_j\right)}. \tag{17}$$

And for S-GsMTLR as

$$\mathrm{RMSE}\left(\mathbf{y}_j, \hat{\mathbf{y}}_j\right) = \sqrt{(1-\frac{1}{n})\left(1 + \mathbf{w}_j^T\hat{\Sigma}_{XX}\mathbf{w}_j - 2\boldsymbol{\beta}_j^T\mathbf{w}_j\right)}, \tag{18}$$

where $\hat{\Sigma}_{XX}$ is the estimated within-covariance used in S-GsMTLR, $\boldsymbol{\beta} \in \mathbb{R}^{d\times c}$ is the regression coefficient or the effect size of univariate GWAS studies used in S-GsMTLR. And we found it quite effective and acceptable in practice.

3.3. Data source

3.3.1. Individual-level neuroimaging genetic data of ADNI

The brain imaging genetic data were obtained from the ADNI database (adni.loni.usc.edu). One primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

The neuroimaging data of the 18-F florbetapir (AV45) PET scans were obtained from the ADNI website, and the demographic information of all subjects were summarized in Table 1. These amyloid imaging data were preprocessed on the basis of the pipeline contained in [32]. After preprocessing, the whole brain were subsampled to generate regions of interest (ROI) measurements based on the MarsBaR automated anatomical labeling (AAL) atlas [33]. For ease of comparison, ten amyloid-derived imaging QTs, which are known to be related to AD, and the details of these imaging QTs were listed in Table 2. Using the regression weights derived from the healthy control participants, these imaging QTs was pre-adjusted to remove the effects of the baseline age, gender, education, and handedness.

Table 2
18-F florbetapir (AV45) PET scans-derived imaging QTs used in this paper.

QT ID	ROI
LHippocampus RHippocampus	Hippocampus
LFrontalMid RFrontalMid	Middle frontal gyrus
LFrontalMedOrb RFrontalMedOrb	Superior frontal gyrus, medial orbital
LFrontalSupMedial RFrontalSupMedial	Superior frontal gyrus, dorsolateral
LRectus RRectus	Gyrus rectus

Table 3
The information of 10 WMM QTs used in S-GsMTLR from the brain white matter microstructure GWAS database.

Name	Information
Average_MD	average value of mean diffusivity across all 21 white matter tracts
Average_RD	average value of radial diffusivity across all 21 white matter tracts
PTR_MD	mean diffusivity of the Posterior thalamic radiation (PTR)
PTR_RD	radial diffusivity of the Posterior thalamic radiation (PTR)
RLIC_MD	mean diffusivity of the Retrolenticular part of internal capsule (RLIC)
RLIC_RD	radial diffusivity of the Retrolenticular part of internal capsule (RLIC)
CGH_MD	mean diffusivity of the Cingulum hippocampus (CGH)
CGH_RD	radial diffusivity of the Cingulum hippocampus (CGH)
SS_AD	axial diffusivity of the Sagittal stratum (SS)
SS_MD	mean diffusivity of the Sagittal stratum (SS)

The genotypic data for the same population used in this study were also downloaded from the ADNI website, which were genotyped by the Human 610-Quad or OmniExpress Array (Illumina, Inc., San Diego, CA, USA). The pre-processing procedure was carried out through the standard quality control (QC) and imputation steps. Secondly, quality-controlled SNPs was calculated by the MaCH software [34] to estimate the missing genotypes. The chromosome 19 sequence contains the well-known AD risk genes such as *APOE, TOMM40*, and *APOC1*. Therefore, we took all SNPs in chromosome 19, including 145,124 SNPs. Our goal is to identify a small portion of SNPs which is correlated with abnormal imaging QTs of AD patients, under the situation where the original imaging genetic data is unavailable.

3.3.2. Summary statistics from GWAS

The brain white matter microstructure (WMM) GWAS was performed to uncover the genetic basis of brain white matters, involving 215 diffusion tensor imaging phenotypes from 34,024 British-ancestry individuals of the UK Biobank (UKB) database [17]. In order to evaluate the performance of our propose method, we extracted 1000*10 and 5000*10 two SNPs-QTs summary statistics matrices from WMM GWAS results in this work. These 1,000 (chr5: 82481553 - 82921104) and 5,000 (chr5: 81947637 - 83988233) SNPs were around the significant locus rs10052710 in chromosome 5, and the detailed information of 10 WMM QTs is provided in Table 3.

The second GWAS we used in this paper was the whole brain imaging-derived phenotypes (IDPs) GWAS database, which studied 8,428 individuals' brain IDPs from the UKB database, covering a comprehensive set of imaging QTs derived from different types of brain imaging data such as diffusion weighted imaging (dMRI) and susceptibility weighted imaging (swMRI or SWI) [18]. We respectively chose two summary statistics matrices of the same sizes from IDP GWAS results. These two matrices encapsulated information of 1,000 (chr5: 82717885 - 82962751) and 5,000 (chr5: 82189046 - 83559329) SNPs around the significant SNP rs13164785 on chromosome 5, and the details of ten IDP QTs is contained in Table 4.

Table 4
The names and its abbreviations of 10 IDP QTs used in S-GsMTLR from the IDPs GWAS database.

Name	Information
rpoicR	IDP_dMRI_TBSS_ICVF_Retrolenticular_part_of_internal _capsule_R
rpoicL	IDP_dMRI_TBSS_ICVF_Retrolenticular_part_of_internal _capsule_L
chR	IDP_dMRI_TBSS_ICVF_Cingulum_hippocampus_R
chL	IDP_dMRI_TBSS_ICVF_Cingulum_hippocampus_L
ccgR	IDP_dMRI_TBSS_ICVF_Cingulum_cingulate_gyrus_R
ccgL	IDP_dMRI_TBSS_ICVF_Cingulum_cingulate_gyrus_L
infR	IDP_dMRI_ProbtrackX_ICVF_ifo_r
infL	IDP_dMRI_ProbtrackX_ICVF_ifo_l
arR	IDP_dMRI_ProbtrackX_ICVF_ar_r
arL	IDP_dMRI_ProbtrackX_ICVF_ar_l

Table 5
RMSE values of S-GsMTR and GsMTR when applied to the ADNI dataset.

	GsMTLR	S-GsMTLR
LHippocampus	0.9880	0.9762
RHippocampus	0.9867	0.9773
LFrontalMedOrb	0.9835	0.9780
RFrontalMedOrb	0.9839	0.9837
LFrontalSupMedial	0.9845	0.9926
RFrontalSupMedial	0.9926	0.9772
LFrontalMid	0.9840	0.9797
RFrontalMid	0.9855	0.9904
LRectus	0.9941	0.9817
RRectus	0.9901	0.9868
mean	0.9873	0.9824

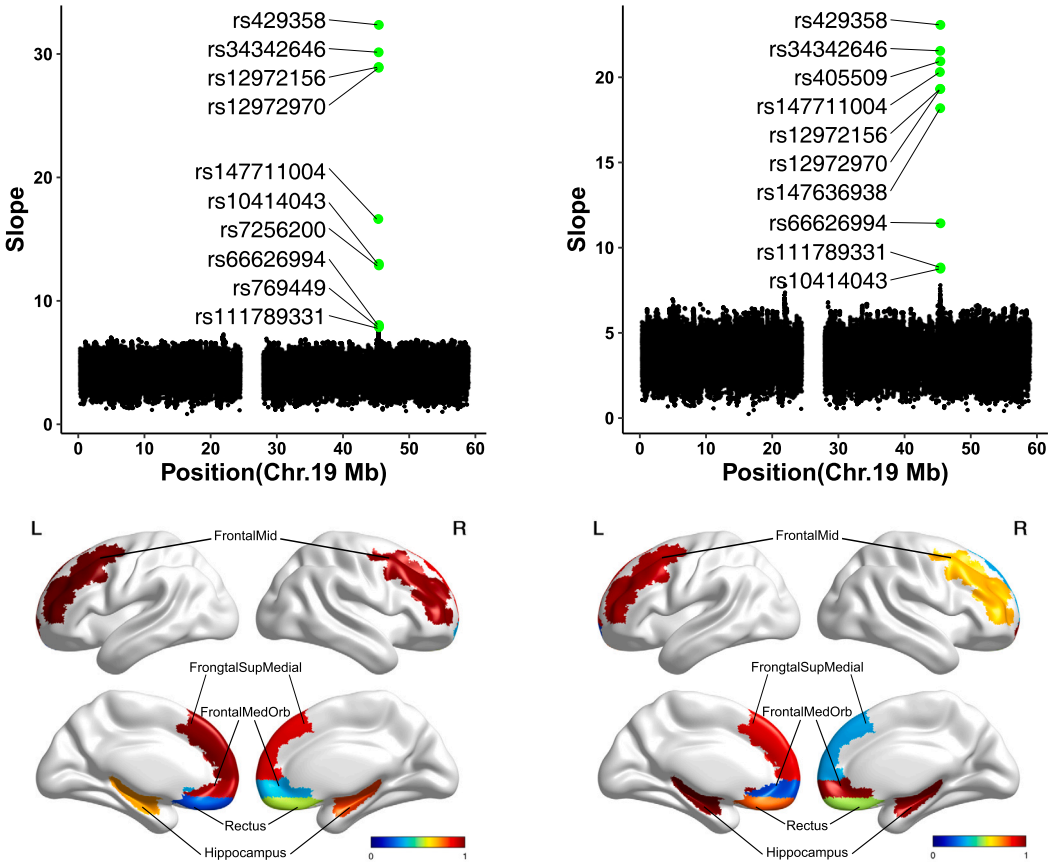
3.4. Study on the ADNI data set

3.4.1. The degree of model fitting

We performed a comparative analysis of the RMSE results for S-GsMTLR and conventional MTLR, as presented in Table 5. The RMSE results of S-GsMTLR were computed by Eq. (18), while those for conventional GsMTLR were calculated by Eq. (17). The results in table illustrated that the RMSE values for S-GsMTLR closely align with those of conventional MTLR, indicating a comparable modeling performance. These results collectively demonstrated that S-GsMTLR achieves a performance akin to that of conventional MTLR.

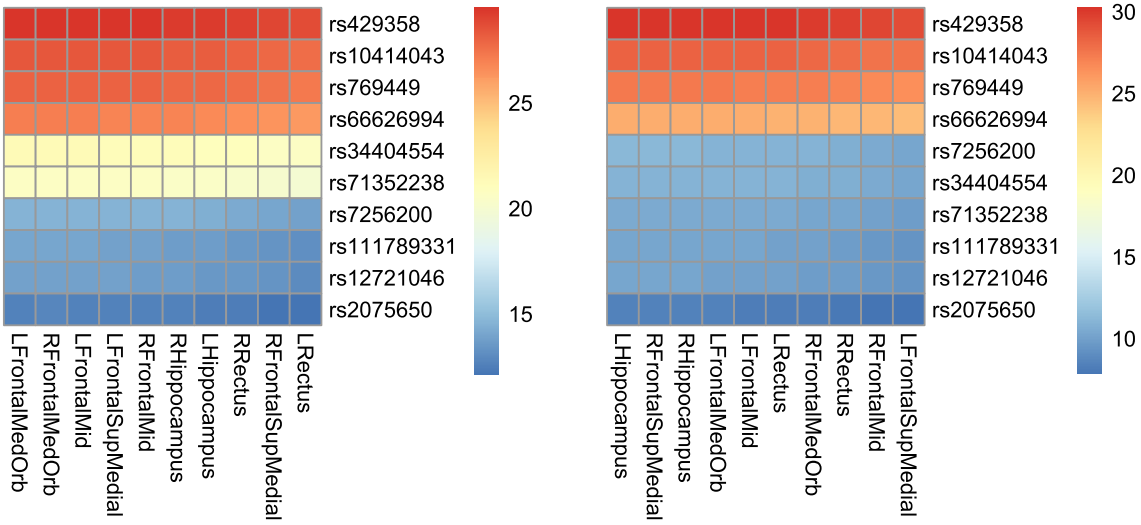
3.4.2. Identification of risk biomarkers

We first compared feature selection results of S-GsMTLR and GsMTLR in terms of SNPs by regression coefficients. Specifically, the scatter plots of the regression weights of SNPs on imaging QTs for both S-GsMTLR and GsMTLR are presented in the top panel of Fig. 1. In each sub-plot, the y-coordinate represents the average effect of each SNP on ten imaging QTs, while the x-coordinate indicates the position of each SNP. A higher scatter point signifies that SNP at this location has a larger effect on QTs. For clarity, we marked and annotated the top ten SNPs in each sub-plot. Both conventional GsMTLR and our S-GsMTLR successfully and the most significant AD risk locus rs429358 (*APOE*). Additionally, all the top ten SNPs identified by both methods were identical and corresponded to well-established AD risk loci, such as rs10414043 (*APOC1*), rs769449 (*APOE*), and rs34404554 (*TOMM40*). This observation indicates that our proposed S-GsMTLR has an equivalent capability to identify genetic factors of AD compared to conventional methods. Furthermore, we also presented the mapping visualization of the average weights of all SNPs on ten imaging QTs in the bottom panel of Fig. 1. To make it clear, we provided the heat map of regression coefficients of top ten SNPs on each imaging QT in Fig. 2. We could clearly observe that



(a) Results for conventional GsMTLR (b) Results for S-GsMTLR

Fig. 1. Weights for SNPs (top panel) and the visualization of 10 imaging QTs (bottom panel) of conventional GsMTLR and S-GsMTLR. We marked and labeled the top 10 SNPs with green color in the top panel. The color coding in bottom panel represents the weights of imaging markers.



(a) Results for conventional MTLR (b) Results for S-GsMTLR

Fig. 2. Top 10 selected SNPs by regression coefficients. The value in each color block is the regression coefficients.

S-GsMTLR and GsMTLR both identified a strong association between SNP rs429358 and all ten imaging QTs. Notably, our method demonstrated that the imaging QTs for the left and right hippocampus had the strongest associations with the top ten genetic variants, indicating superior performance of the proposed method.

We also quantitatively compared the regression weights between S-GsMTLR and conventional GsMTLR. The correlation coefficient, denoted as ρ, ranges from -1 to 1, with higher absolute values indicating closer equivalence and zero indicating no relationship. All ρ values for the ten imaging QTs were significantly less than 1 ($p < 0.0001$), indicating

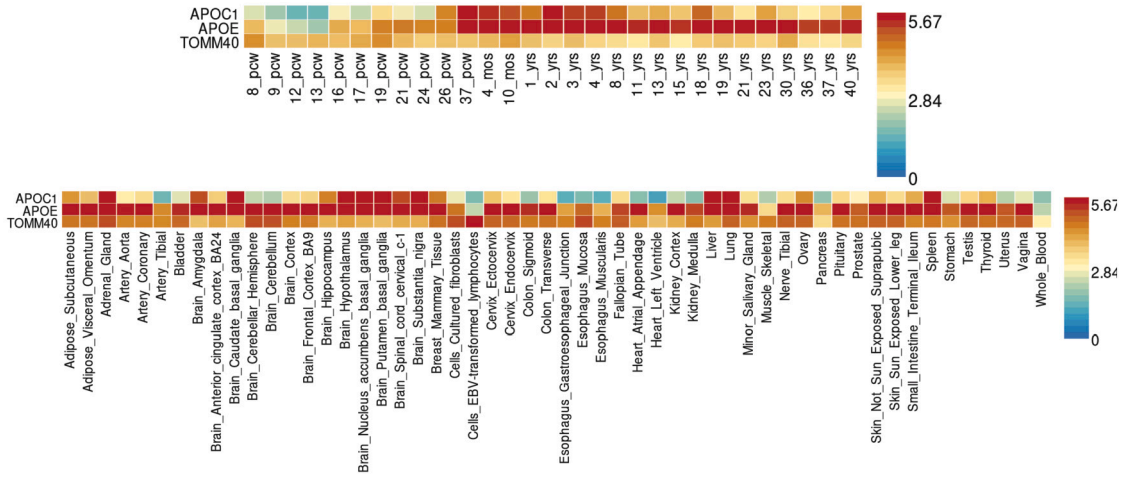


Fig. 3. Heat maps of normalized gene expression value (zero mean normalization of log2 transformed expression) for prioritized genes for top ten SNPs, for GTEx v8 RNAseq data (bottom panel) and BrainSpan data (top panel). The letters in the bottom panel label represent time units for life stages, pcw for post-conception weeks, mos for months, and yrs for years.

Table 6
Top ten SNPs selected by both conventional and our proposed method.

GsMTLR		S-GsMTLR	
SNPs	*p*-value	SNPs	*p*-value
rs429358	2.35E-16	rs429358	2.35E-16
rs10414043	4.60E-13	rs10414043	4.60E-13
rs769449	3.28E-13	rs769449	3.28E-13
rs66626994	1.07E-08	rs66626994	1.07E-08
rs34404554	2.47E-08	rs7256200	4.60E-13
rs71352238	4.07E-08	rs34404554	2.47E-08
rs7256200	4.60E-13	rs71352238	4.07E-08
rs111789331	2.15E-09	rs111789331	2.15E-09
rs12721046	1.49E-09	rs12721046	1.49E-09
rs2075650	2.47E-08	rs2075650	2.47E-08

that S-GsMTLR has a strong agreement with conventional GsMTLR in identifying relevant biomarkers.

In summary, the above results consistently indicated that S-GsMTLR possesses feature selection capabilities comparable to conventional GsMTLR. Notably, S-GsMTLR demonstrated a more robust performance in the regression weights in terms of imaging QTs. Consequently, S-GsMTLR emerges as a compelling and robust approach for identifying the genetic underpinnings of interested imaging QTs, eliminating the need for access to the original individual-level imaging genetic data.

3.4.3. Effectiveness of genetic risk factors

To assess the statistical significance of the identified genetic variants, we applied a one-way analysis of variance (ANOVA) to evaluate the impact of the top ten SNPs on diagnosis status, with age, sex, handedness, and years of education as covariates. A genetic variant was considered significant if its primary effect on diagnostic status was statistically significant. The *p*-values from the one-way ANOVA for the top ten SNPs identified by our method, as well as by the conventional GsMTLR, are summarized in Table 6. All *p*-values were statistically significant ($p < 0.05$), demonstrating that our method can effectively identify AD risk variants by leveraging summary statistics from GWAS.

3.4.4. Gene expression analyses

To further evaluate the biological significance of the identified loci at the gene expression level, we utilized GENE2FUNC tool of FUMA [35] for gene expression analysis. This tool enables us to explore gene expression patterns associated with the top ten identified SNPs. We incorporated data from the GTEx database (Version 8), which includes 54 tissue types, as well as BrainSpan RNA sequencing data covering 29 developmental stages. Using these datasets, we constructed heat maps of gene expressions, with each heat map representing the average normalized expression value for its respective label.

We presented mRNA expression profiles of priority genes associated with the top 10 SNPs on chromosomes 19 across 54 developing and adult brain tissue types, as shown in Fig. 3. The bottom panel presents a heat map of gene expression based on GTEx version 8 RNA sequencing data, highlighting expression levels in various brain tissues for the genes *APOE*, *APOC1*, and *TOMM40*. In the BrainSpan data, these genes exhibited high expression levels throughout life cycle as shown in the top panel of Fig. 3. Specifically, *APOE* and *TOMM40* remains most highly expression across all life stages, while *APOC1* shows increased expression in the late prenatal stages (26 post-conception weeks, 26_pcw). These findings underscore the effectiveness of our S-GsMTLR method in identifying genetic variants associated with various human brain tissues across different stages of the life cycle.

All these results not only demonstrated the powerful feasibility of S-GsMTLR, but also its effectiveness in a sparse linear regression model for imaging genetic studies without requiring individual-level data.

3.5. Application to summary statistics from brain imaging GWAS

Our goal is to use the proposed S-GsMTLR method to conduct group-sparse multivariate multi-task analysis on summary statistics from large GWAS, identifying meaningful associations and genetic variations without the need for original imaging genetic data.

3.5.1. Application to summary statistics from the brain WMM GWAS

To evaluate the model-fitting capability of our proposed S-GsMTLR when applied to GWAS summary statistics without access to the original individual-level imaging and genetic data, we first present the RMSE values for datasets containing 1,000 and 5,000 SNPs in Table 7. The results indicate that the RMSE values for all ten DTI QTs were low, demonstrating the robust modeling capability of the S-GsMTLR method.

The scatter plots of the regression coefficients are displayed in Fig. 4, with the top ten genetic variants marked and labeled. Both sub-plots revealed the same top ten SNPs for datasets containing 1,000 and 5,000 SNPs, demonstrating the scalability and stability of S-GsMTLR. For clarity, we also presented the regression weights of the top ten SNPs on each imaging QT in Fig. 5. In this figure, rs10052710 (*VCAN*) exhibite the strongest correlation with all ten QTs, and the association between this locus and the imaging QT Average-MD (average value of mean diffusivity across 21 white matter tracts) has the highest coefficient value. Additionally, we present the locus plot of the lead SNP rs10052710 in

Table 7
RMSE values of ten QTs and its average when applied S-GsMTLR to the summary statistics from brain white matter microstructure GWAS for data set with 1,000 SNPs and 5,000 SNPs.

Date Set	Average_MD	Average_RD	PTR_MD	PTR_RD	CGH_MD	CGH-RD	RLIC-MD	RLIC_RD	SS_MD	SS_AD	Mean
1,000 SNPs	0.9248	0.9297	0.9351	0.9363	0.9418	0.9546	0.9428	0.9464	0.9275	0.9621	0.9401
5,000 SNPs	0.9889	0.9896	0.9903	0.9905	0.9913	0.9932	0.9914	0.9920	0.9893	0.9943	0.9911

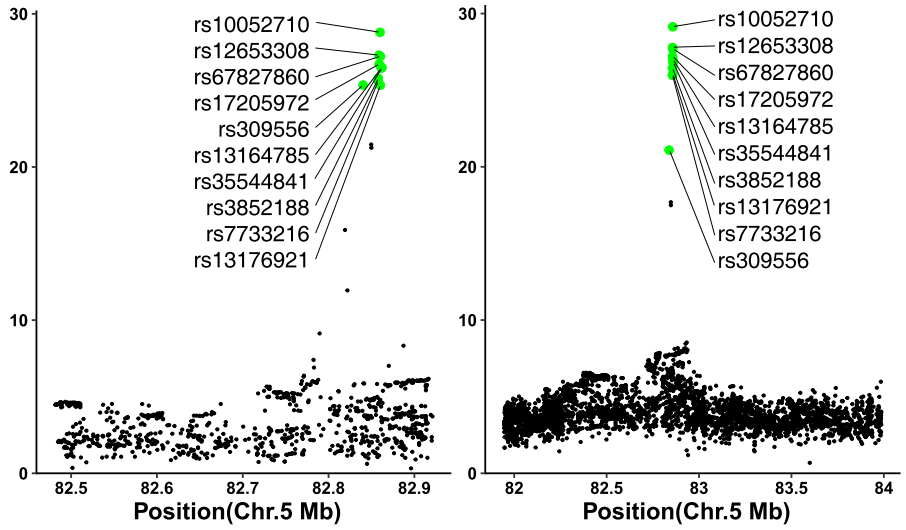
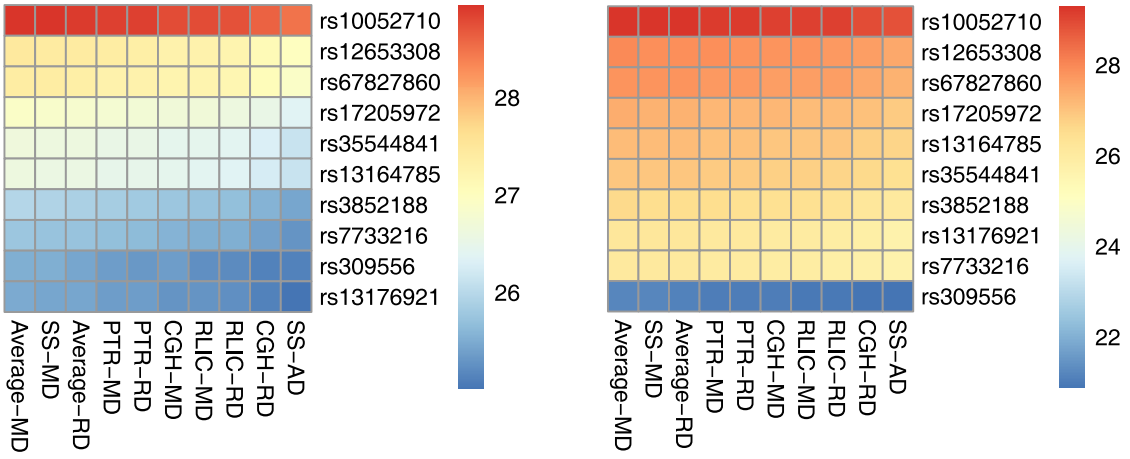


Fig. 4. Regression weights when applied S-GsMTR to brain white matter microstructure GWAS. The left and right panel presented the results for data set with 1,000 and 5,000 SNPs respectively.



(a) Results for data set with 1,000 SNPs

(b) Results for data set with 5,000 SNPs

Fig. 5. The top ten selected SNPs by regression coefficients when applied S-GsMTLR to the GWAS summary statistics from brain white matter microstructure GWAS.

Fig. 6, which showed that rs10052710 is an intron variant of *VCAN* on chromosome 5, exhibiting the highest significance level linked to all ten imaging biomarkers. These findings suggest that rs10052710 might be primarily responsible for brain white matter microstructural differences and abnormalities, warranting further investigation. Furthermore, both the scatter plots and heat maps clearly depicted the group structure of the top ten identified risk variants. For example, rs12653305 (*VCAN*) and rs67827860 (*VCAN*) showed similar patterns in both the scatter plots and heat maps. Similar results were observed for rs17205972 (*VCAN*), rs35544841 (*VCAN*), and rs13164785 (*VCAN*).

These findings demonstrate that S-GsMTLR not only can identify loci reported by GWAS but also excels in sparse feature selection and reveal group structures for multiple SNPs within the same gene. This highlights the superiority of S-GsMTLR over single-variable GWAS in structural information mining.

Moreover, we present a comprehensive overview of the gene expression of the top ten SNPs in Fig. 7. Notably, the top ten SNPs identified by S-GsMTLR are all linked to the gene *VCAN*. The heat map presented in the image above vividly illustrates the expression levels of the *VCAN* gene in different brain tissues. In the bottom panel, we can observe that *VCAN* expression is highest during the early prenatal period (9-24 weeks after conception), while the reverse is true during the postpartum period. In summary, these two heat maps show that our methods have successfully identified the genetic basis of brain tissue with elevated expression levels during human brain development.

Furthermore, Fig. 7 provides a comprehensive overview of gene expression for the top ten SNPs. Notably, these SNPs, identified by S-GsMTLR, are all linked to the *VCAN* gene. In the top panel, *VCAN* expression is highest during the early prenatal period (8-24 weeks after conception) and decreases during the postpartum period. The heat map

Table 8
RMSE values of ten QTs and its average value when applied S-GsMTLR to the summary statistics from brain IDPs GWAS for data set with 1,000 SNPs and 5,000 SNPs.

Data Set	Rrpoic	Lrpoic	Rch	Lch	Rccg	Lccg	Rinf	Linf	Rar	Lar	Mean
1,000 SNPs	0.9776	0.9779	0.9731	0.9741	0.9723	0.9767	0.9726	0.9737	0.9757	0.9729	0.9747
5,000 SNPs	0.9980	0.9980	0.9976	0.9977	0.9976	0.9980	0.9976	0.9977	0.9979	0.9976	0.9978

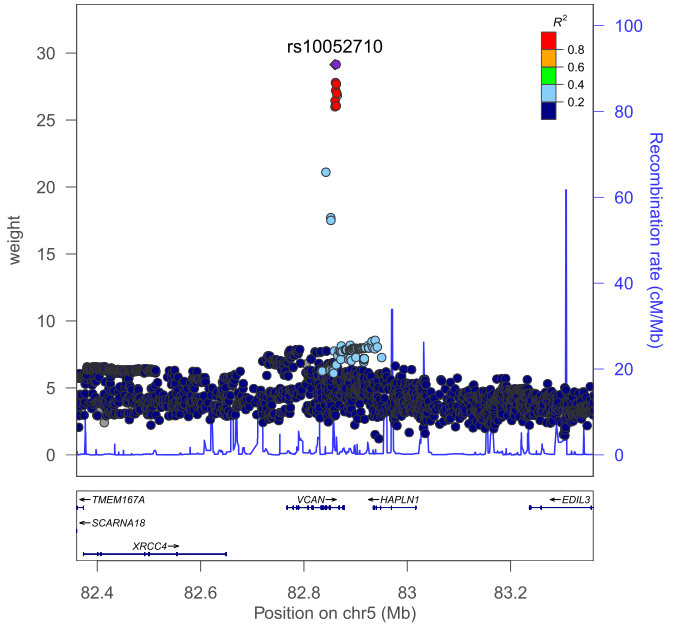


Fig. 6. Locus plots for the top lead SNP rs10052710.

in the bottom panel vividly illustrates the expression levels of *VCAN* in various brain tissues. These heat maps collectively demonstrate that our methods have successfully identified genetic variants associated with elevated gene expression in brain tissues during human brain development.

Taken together, all above results demonstrated that S-GsMTLR can not only model the associations between multiple genetic variations and multiple brain imaging DTIs from summary statistics, but also successfully discovered important genetic variations being responsible for brain white matter microstructures.

3.5.2. *Application to summary statistics from the comprehensive IDP GWAS*

We further investigated the performance of S-GsMTLR using summary statistics from another brain imaging GWAS database. The RMSE results, summarized in Table 8. We can observe that the RMSE values for all imaging QTs were consistently low, demonstrating our method's powerful multivariate multi-task modeling capability.

Fig. 8 shows the average regression weights of SNPs across all ten imaging QTs, with significant loci highlighted and annotated. For clarity, Fig. 9 presents the regression weights of the top ten SNPs for each imaging QTs. Notably, rs13164785 (*VCAN*) exhibited the strongest relationships with all ten imaging QTs, consistent with the results of univariate GWAS. Importantly, both the scatter plots and heat maps indicate that S-GsMTLR identified group structures among several SNPs, which were overlooked by GWAS. For instance, the variants rs13164785 (*VCAN*) and rs67827860 (*VCAN*) had nearly identical weights, suggesting similar or comparable functionality. Interestingly, both rs13164785 and rs67827860 belong to the *VCAN* gene, with LD scores of 1 [36].

Most importantly, S-GsMTLR identified the GWAS-missed locus rs309587 (*VCAN*), which was later reported by the authors in an expanded IDP GWAS study [37] and confirmed by other researchers [38]. In Fig. 10, the locus plot of rs309587 illustrated that this lead SNP is an intronic variant situated within the *VCAN* gene on chromosome 5. Then we performed a phenome-wide association study (pheWAS) analysis for this locus to investigate its potential associations with diverse array of phenotypes across 28 domains, as illustrated in Fig. 11. pheWAS was performed using publicly available data from the GWAS Atlas [39] (https://atlas.ctglab.nl). The figure revealed a noteworthy association of rs309587 with neurological phenotypes, as well as metabolic and skeletal traits. Consequently, this feature selection results demonstrated that S-GsMTLR can not only replicate the GWAS results quite well, but also outperform it in terms of genetic basic identification and the structure information identification. In summary, these results demonstrated that S-GsMTLR performed quite well in brain imaging genetic studies by only using the summary statistics from GWAS.

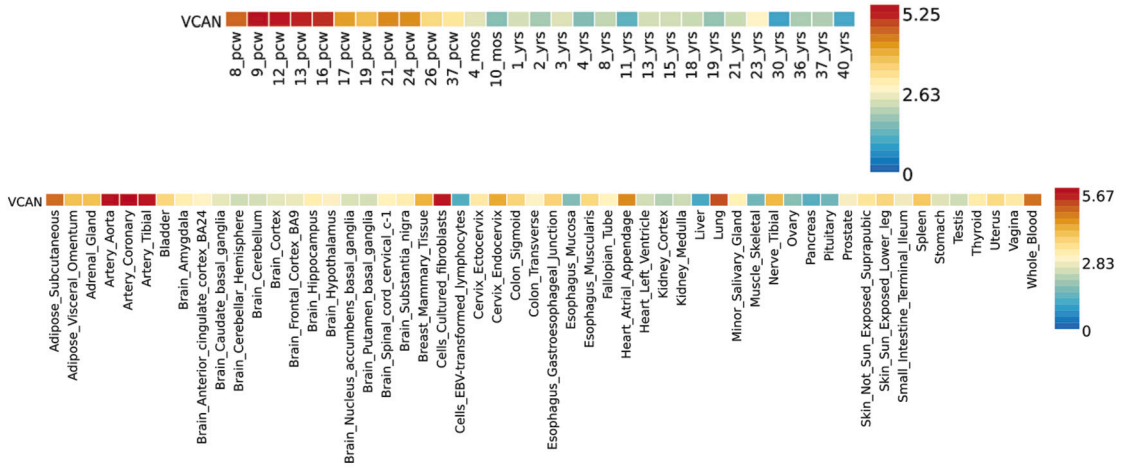


Fig. 7. Heat maps of normalized gene expression value (zero mean normalization of log2 transformed expression) for prioritized genes for top ten SNPs, for BrainSpan data (bottom panel) and GTEx v8 RNAseq data (top panel). The letters in the bottom panel label represent time units for life stages, pcw for post-conception weeks, mos for months, and yrs for years.

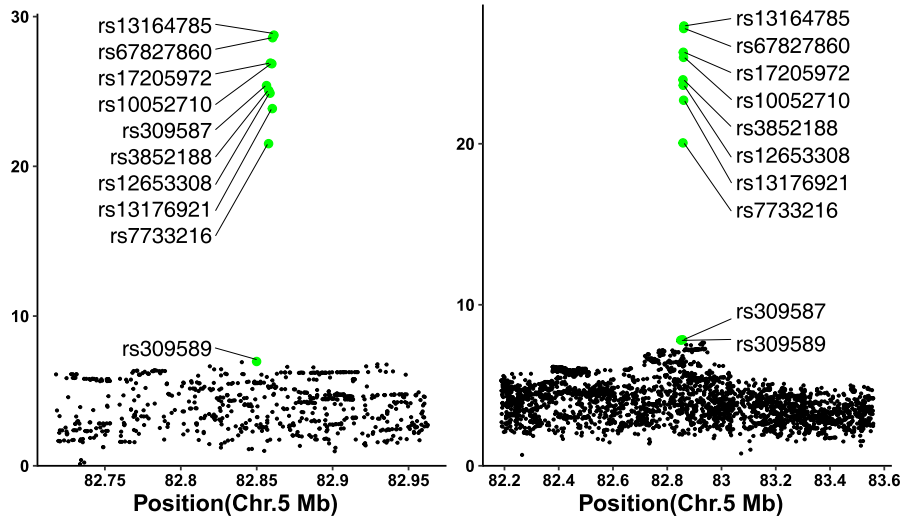
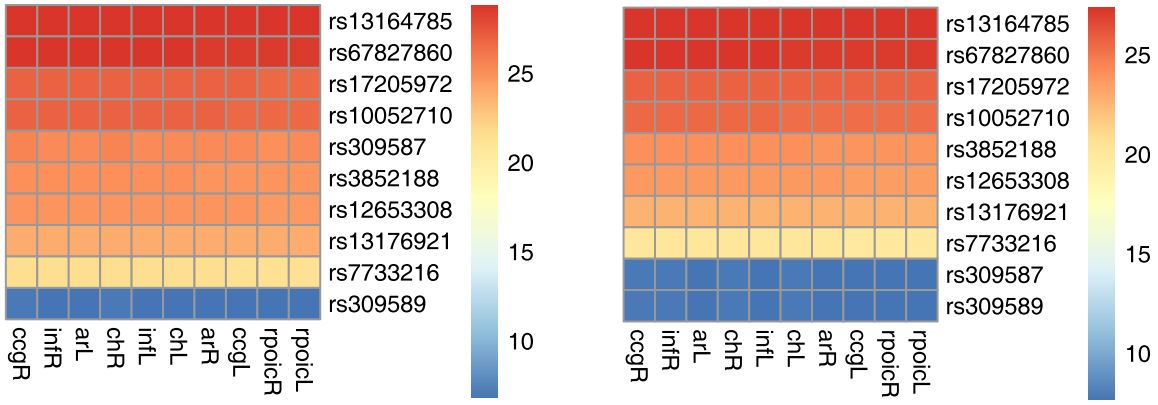


Fig. 8. Regression weights when applied S-GsMTR to brain IDPs GWAS. The left and right panel presented the results for data set with 1,000 and 5,000 SNPs respectively.



(a) Results for data set with 1,000 SNPs

(b) Results for data set with 5,000 SNPs

Fig. 9. The top 10 selected SNPs by regression coefficients when applied S-GsMTLR to the summary statistics from brain IDPs GWAS.

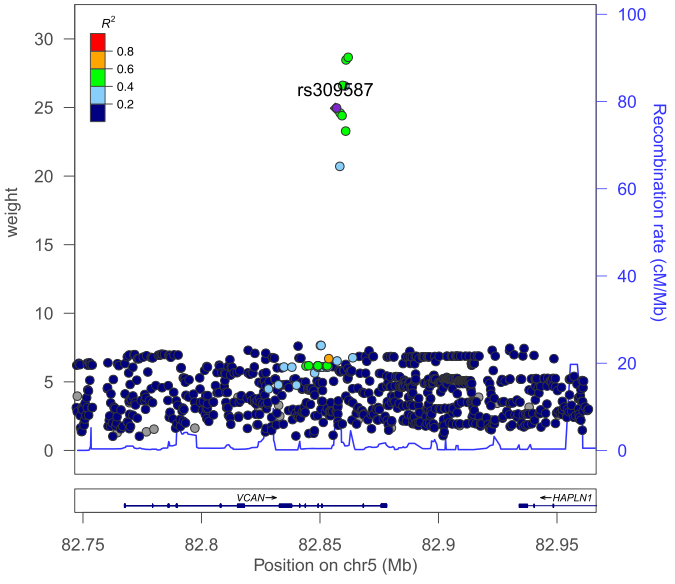


Fig. 10. Locus plots for SNP rs309587.

4. Discussion and conclusion

Conventional sparse multi-task learning methods have played pivotal roles in brain imaging genetics [40,41,11]. However, these methods rely on individual-level imaging and genetic data, which limits their broader application. Meanwhile, publicly available GWAS results typically identify associations between single SNP and single QT, potentially overlooking meaningful information related to multiple variations and multiple traits. To address these limitations, we proposed S-GsMTLR (group-sparse multivariate multi-task linear regression method based on summary statistics from GWAS) for brain imaging genetics. S-GsMTLR applies multivariate multi-task analysis on univariate GWAS results, aiming to study genetic basic of interested multiple imaging QTs while not accessing the individual-level data. We have proved that this strategy was reasonable and practical being supported by Theorem 1. Results on ADNI database and two additional GWAS databases showed that S-GsMTLR performed quite well without accessing the original individual-level imaging and genetic data. The model's capability to utilize preselected imaging QTs for identifying relevant genetic variants even outperformed than GWAS. In practice, our method could obtain comparable results to conventional one if two conditions were satisfied. First, the reference population was appropriately chosen, which guaranteed the correctness of the covariance information. Second, the number

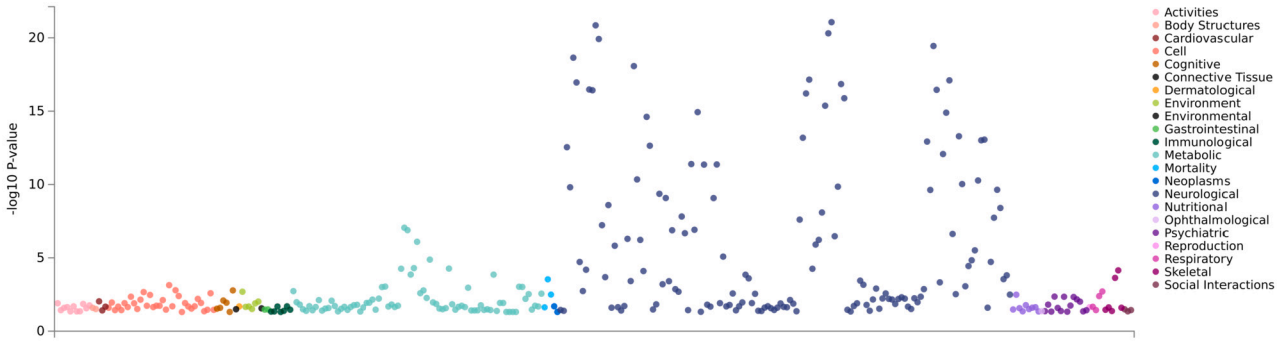


Fig. 11. pheWAS result for rs309587. PheWAS plot presents the significance of rs309587 on a range of traits based on MAGMA gene-based tests (Bonferroni corrected *P*-value threshold: 7.51e-7).

of the reference population should not be small which guaranteed the covariance's goodness of fit. Since 1kGP database provide diverse ethnic groups and enough subjects, these two conditions could be met in most cases [42,43].

All in all, our group-sparse multivariate multi-task learning method, S-GsMTLR, proved to be an effective and powerful computational strategy in the realm of brain imaging genetic studies. It is worth emphasizing that our approach is not bound by a specific regression model, instead, it can accommodate a wide range of regression methods. Moving forward, it is imperative to incorporate pathway and brain network information into our sparse learning framework.

Fundings

This work was supported in part by the STI2030-Major Projects (2022ZD0213700); National Natural Science Foundation of China [61936007, 62136004, U23A20335, 62373306]; Fundamental Research Funds for the Central Universities; Natural Science Basic Research Program of Shaanxi [2020JM-142]; and China Postdoctoral Science Foundation [2020T130537] at Northwestern Polytechnical University.

CRediT authorship contribution statement

Duo Xi: Writing – original draft, Software, Methodology. **Dingnan Cui:** Resources, Investigation. **Mingjianan Zhang:** Investigation. **Jin Zhang:** Writing – review & editing, Validation. **Muheng Shang:** Writing – review & editing. **Lei Guo:** Conceptualization. **Junwei Han:** Writing – review & editing, Supervision, Funding acquisition. **Lei Du:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare 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. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; JUnjunnjun Fujirebio; GE Healthcare; IXICO Ltd.; 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.

Data used in preparation of this article were obtained from the 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/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

[1] Du L, Liu K, Zhu L, Yao X, Risacher SL, Guo L, et al. Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: a longitudinal study of the ADNI cohort. Bioinformatics 2019;35:i474–83.
[2] Li G, Han D, Wang C, Hu W, Calhoun VD, Wang Y-P. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. Comput Methods Programs Biomed 2020;183:105073.
[3] Thompson PM, Martin NG, Wright MJ. Imaging genomics. Curr Opin Neurol 2010;23:368–73.
[4] Shen L, Thompson PM. Brain imaging genomics: integrated analysis and machine learning. Proc IEEE 2019;108:125–62.
[5] Wen C, Ba H, Pan W, Huang M, Initiative ADN. Co-sparse reduced-rank regression for association analysis between imaging phenotypes and genetic variants. Bioinformatics 2020;36:5214–22.
[6] Silver M, Janousova E, Hua X, Thompson PM, Montana G, Initiative ADN, et al. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. NeuroImage 2012;63:1681–94.
[7] Zhu Xiaofeng, Suk Heung-Il, Huang Heng, Shen Dinggang. Structured sparse low-rank regression model for brain-wide and genome-wide associations. In: Medical image computing and computer-assisted intervention: MICCAI. Springer; 2016. p. 344–52.
[8] Wang M, Shao W, Hao X, Zhang D. Identify complex imaging genetic patterns via fusion self-expressive network analysis. IEEE Trans Med Imaging 2021;40:1673–86.
[9] Silver M, Montana G, Initiative ADN, et al. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat Appl Genet Mol Biol 2012;11:000010220215446115 1755.
[10] Wang Hua, Nie Feiping, Huang Heng, Kim Sungeun, Nho Kwangsik, Risacher Shannon L, 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.
[11] Huang M, Chen X, Yu Y, Lai H, Feng Q. Imaging genetics study based on a temporal group sparse regression and additive model for biomarker detection of Alzheimer's disease. IEEE Trans Med Imaging 2021;40:1461–73.
[12] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017;101:5–22.

[13] Xi D, Cui D, Zhang J, Shang M, Zhang M, Guo L, et al. Identification of disease-sensitive brain imaging phenotypes and genetic factors using GWAS summary statistics. In: Medical image computing and computer assisted intervention – MICCAI. Springer; 2023. p. 622–31.
[14] Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature 2017;542:186–90.
[15] Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 2013;45:1452–8.
[16] 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.
[17] Zhao B, Li T, Yang Y, Wang X, Luo T, Shan Y, et al. Common genetic variation influencing human white matter microstructure. Science 2021;372:eabf3736.
[18] Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 2018;562:210–6.
[19] Apostolova LG, Risacher SL, Duran T, Stage EC, Goukasian N, West JD, et al. Associations of the top 20 Alzheimer disease risk variants with brain amyloidosis. JAMA Neurol 2018;75:328–41.
[20] 1000 Genomes Project Consortium, et al. A map of human genome variation from population scale sequencing. Nature 2010;467:1061.
[21] Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics 2016;32:1981–9.
[22] Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet 2018;50:229–37.
[23] Gai L, Eskin E. Finding associated variants in genome-wide association studies on multiple traits. Bioinformatics 2018;34:i467–74.
[24] 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.
[25] Luo L, Shen J, Zhang H, Chhibber A, Mehrotra DV, Tang Z-Z. Multi-trait analysis of rare-variant association summary statistics using MTAR. Nat Commun 2020;11:2850.
[26] Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369–75.
[27] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015;4. s13742–015.
[28] Zhao Z, Yi Y, Song J, Wu Y, Zhong X, Lin Y, et al. PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. Genome Biol 2021;22:1–19.
[29] Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat Commun 2021;12:4192.
[30] Zhao Z, Fritsche LG, Smith JA, Mukherjee B, Lee S. The construction of cross-population polygenic risk scores using transfer learning. Am J Hum Genet 2022;109:1998–2008.
[31] 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. Nature 2015;526:68.
[32] Ashburner J, Friston KJ. Voxel-based morphometry—the methods. NeuroImage 2000;11:805–21.
[33] 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.
[34] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–34.
[35] Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017;8:1826.
[36] Rutten-Jacobs LC, Tozer DJ, Duering M, Malik R, Dichgans M, Markus HS, et al. Genetic study of white matter integrity in UK Biobank (N= 8448) and the overlap with stroke, depression, and dementia. Stroke 2018;49:1340–7.
[37] Smith SM, Douaud G, Chen W, Hanayik T, Alfaro-Almagro F, Sharp K, et al. An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. Nat Neurosci 2021;24:737–45.
[38] Zhao B, Zhang J, Ibrahim JG, Luo T, Santelli RC, Li Y, et al. Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n= 17,706). Mol Psychiatry 2021;26:3943–55.
[39] Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJ, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet 2019;51:1339–48.
[40] Wang Meiling, Hao Xiaoke, Huang Jiashuang, Shao Wei, Zhang Daoqiang. Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease. Bioinformatics 2018;35:1948–57.
[41] Zhou Tao, Thung Kim-Han, Liu Mingxia, Shen Dinggang. Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. IEEE Trans Biomed Eng 2019;66:165–75.
[42] Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 2018;9:1825.
[43] Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet 2021;53:1276–82.