Published in final edited form as:
Can J Stat. 2021 March ; 49(1): 182–202. doi:10.1002/cjs.11605.

A Structured Brain-wide and Genome-wide Association Study Using ADNI PET Images

Yanming Li[a], **Bin Nan**[b,*], **Ji Zhu**[c], **Alzheimer's Disease Neuroimaging Initiative**

[a]Department of Biotatistics & Data Science, University of Kansas Medical Center Kansas City, KS 66160

[b]Department of Statistics, University of California at Irvine Irvine, CA 92697

[c]Department of Statistics, University of Michigan Ann Arbor, MI 48109

Abstract

A multi-stage variable selection method is introduced for detecting association signals in structured brain-wide and genome-wide association studies (brain-GWAS). Compared to conventional single-voxel-to-single-SNP approaches, our approach is more efficient and powerful in selecting the important signals by integrating anatomic and gene grouping structures in the brain and the genome, respectively. It avoids large number of multiple comparisons while effectively controls the false discoveries. Validity of the proposed approach is demonstrated by both theoretical investigation and numerical simulations. We apply the proposed method to a brain-GWAS using ADNI PET imaging and genomic data. We confirm previously reported association signals and also find several novel SNPs and genes that either are associated with brain glucose metabolism or have their association significantly modified by Alzheimer's disease status.

Keywords

Brain-wide and genome-wide association studies; multivariate sparse group lasso (MSGLasso); structured high-dimensional multivariate linear regression; ultrahigh-dimensional predictors; ultrahigh-dimensional responses

1 INTRODUCTION

Human brain structures are highly heritable [6, 41]. The association patterns between the brain and the genome offer important information about development and progression mechanisms of chronic cognitive diseases such as Alzheimer's disease (AD) [34]. Modern technologies of neuroimaging scan and next generation sequencing enable us to look at such association patterns at the resolutions of single voxel and single-nucleotide polymorphism (SNP) scales. However, given the enormous numbers of variables in both imaging data (~ millions of voxels) and genotype data (~ millions of SNPs), it is extremely challenging

[*] bnan@uci.edu .

SUPPORTING INFORMATION
Additional results from analyzing the ADNI PET imaging and genetic data are provided in the online Supporting Information.

to detect the true association signals immersed in the ultrahigh-dimensional noises. Many current brain-wide and genome-wide association studies (brain-GWAS) look at a single-voxel-to-single-SNP pair at a time [46]. Such single-voxel-to-single-SNP (or pairwise) approaches suffer from very limited power in detecting the true signals, mostly due to the astronomical number of multiple comparisons needed to control the false positive discoveries [46, 19].

Marginal pairwise approaches treat different voxel-to-SNP pairs as independent. A joint model with all voxels and all SNPs considered together is often of more scientific interest. Compared to marginal pairwise approaches, joint modeling has enormous potential to improve the power of detecting association signals. The multivariate linear regression is a common way for jointly modeling multiple responses and multiple predictors. However, such a model is ill-posed when the dimensions of responses and predictors are both greater than the sample size, as the solution is not unique. Another limitation of marginal pairwise approaches is that they fail to incorporate the intrinsic biological grouping structures, such as anatomical regions of interest (ROI) in the brain and genes in the genome, respectively. Figure 1 illustrates an atlas of anatomical ROIs and their positions in the brain.

Li et al. [29] introduced a multivariate sparse group lasso (MSGLasso), a regularization method for high-dimensional multivariate-response and multiple-predictor linear regression with grouping structures on both responses and predictors. They show that the power of detecting the true association signals can be significantly increased by incorporating the grouping structures. However, it is computationally infeasible to directly fit the MSGLasso with ultrahigh-dimensional neuroimaging and genomic data, where the numbers of responses and predictors are of exponential orders of the sample size. As in our brain-GWAS, each response image consists of $Q \approx 350,000$ voxels and each genome consists of $P \approx 560,000$ SNPs, while we only have $n = 373$ samples. Furthermore, conditions that guarantee selection consistency for the MSGLasso may fail to hold for ultrahigh-dimensional cases [28].

To address these challenges, we propose a multi-stage variable selection method for settings with ultrahigh-dimensional responses and ultrahigh-dimensional predictors, both with grouping structures. The proposed method consists of two selection stages. The first selection stage aims to remove unimportant response-to-predictor group pairs. The second stage then selects important individual-level signals only within the selected group pairs. Stability selection [35] is used in both stages to enhance the stability of the selection and control false positives.

The contribution of the proposed method to variable selection is two-fold. Firstly, it is a joint modeling approach with both ultrahigh-dimensional responses and ultrahigh-dimensional predictors. It avoids the huge number of downstream hypothesis tests and multiple comparisons. Secondly, it is a structured approach taking into consideration the grouping structures of both responses and predictors. These unique characteristics enable the proposed method to significantly increase the power of identifying true signals, and at the same time, reduce the number of false discoveries.

The proposed method is particularly useful in conducting structured brain-wide and genome-wide association studies (brain-GWAS). In this article, we applied it to Fluorine-uorodeoxiglucose positron emission tomography (FDG-PET) neuroimaging data and DNA genotyping data collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database for detecting association signals between voxel-level neuroimaging phnotypes and genetic variants. FDG-PET images measure brain glucose metabolism, and can reflect changes of brain metabolic pattern as diagnostics of AD progression [37]. We emphasize that the proposed method is applicable to a wide range of brain-GWASs with different imaging modalities or molecular data types, such as functional magnetic resonance imaging (fMRI), methylation, copy number variation and mitochondria DNA profiles, or with different grouping structures, such as neuroimages grouped by functional regions, cortices and genomic profiles grouped by gene pathways, protein networks, etc. To the best of our knowledge, our work is the first one to conduct a structured brain-GWAS at voxel and SNP levels using a joint model. Compared to the pairwise approaches [46, 19] and other marginal approaches such as gene based analysis [22] that regresses each single voxel on a set of SNPs within a gene, our approach is able to identify more genetic signals that either are associated with brain glucose metabolism or have their association significantly modified by AD status. Computationally, the proposed method is in general more efficient compared to the pairwise approaches [46]. The major computational cost saving comes from the dimension reduction in the first selection stage and the fact that we only focus on the selected ROI-to-gene pairs in the downstream analyses.

2 Model and method

Details of our proposed model and method are provided in this Section as the background of conducting a structured brain-GWAS for the ADNI PET imaging and genomic data. The main procedure consists of two selection stages in a multivariate linear regression model with the ultimate goal being to efficiently and jointly select the important association signals between ultrahigh-dimensional neuroimaging responses and genetic DNA predictors.

Let $\boldsymbol{Y}$ be the $n \times Q$ matrix of voxel-level neuroimaging-responses, $\boldsymbol{X}$ be the $n \times P$ matrix of SNP genotypes. We consider the following multivariate linear regression model

$$\begin{aligned} Y = \mathbf{I}\beta_0^{\mathrm{T}} + XB_X + \mathbf{I}_{ad}\beta_{ad}^{\mathrm{T}} + \mathbf{I}_{mci}\beta_{mci}^{\mathrm{T}} + (X \times \mathbf{I}_{ad})B_{Xad} + (X \times \mathbf{I}_{mci})B_{Xmci} + \\ \mathrm{Age}\beta_{age}^{\mathrm{T}} + Sex\beta_{sex}^{\mathrm{T}} + E, \end{aligned} \tag{1}$$

where $\mathbf{I}$ is a length-n vector with entries 1, $\mathbf{I}_{ad}$ and $\mathbf{I}_{mci}$ are length-n indicators for AD and mild cognitive impairment (MCI) subjects, respectively, $\boldsymbol{X} \times \mathbf{I}_{ad}$ and $\boldsymbol{X} \times \mathbf{I}_{mci}$ are $n \times P$ matrices of interaction terms between genetic predictors and disease status, **Age** and **Sex** are length-n covariate vectors of age and sex, respectively. Here $\boldsymbol{\beta}_0$ is a length-Q grand intercept vector; $\boldsymbol{\beta}_{ad} = \mathbf{I}_Q\beta_{ad}$, $\boldsymbol{\beta}_{mci} = \mathbf{I}_Q\beta_{mci}$, $\boldsymbol{\beta}_{age} = \mathbf{I}_Q\beta_{age}$, $\boldsymbol{\beta}_{sex} = \mathbf{I}_Q\beta_{sex}$ are coefficient vectors for AD indicator, MCI indicator, age and sex, respectively, where $\mathbf{I}_Q$ is a length-Q vector with entries 1; $\boldsymbol{B}$, $\boldsymbol{B}_{Xad}$, $\boldsymbol{B}_{Xmci}$ are regression coefficient matrices for genetic, genetic-AD interaction and genetic-MCI interaction effects, respectively; $\mathbf{E}$ is an $n \times Q$ matrix of noise terms arising from a Q-dimensional multivariate normal distribution with zero means. The superscript T represents transpose of a matrix or vector.

When variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered, $\boldsymbol{\beta}_0$ is zero and model (1) reduces to

$$Y = \mathbb{X}\mathbb{B} + \mathbf{E} \tag{2}$$

with $\mathbb{X} = (X, \mathbf{I}_{ad}, \mathbf{I}_{mci}, X \times \mathbf{I}_{ad}, X \times \mathbf{I}_{mci}, \mathbf{Age}, \mathbf{Sex})$ being the grand predictor matrix and $\mathbb{B} = \left(\boldsymbol{B}_X^{\mathrm{T}}, \beta_{ad}, \beta_{mci}, \boldsymbol{B}_{Xad}^{\mathrm{T}}, \boldsymbol{B}_{Xmci}^{\mathrm{T}}, \beta_{age}, \beta_{sex}\right)^{\mathrm{T}}$ being the grand coefficient matrix. Here we do not require the selection to respect the model hierarchy, i.e., an interaction term can be selected into the final model even if the baseline genetic main effect is not selected.

When the imaging responses $\boldsymbol{Y}$ and genetic predictors $\boldsymbol{X}$ are grouped into ROIs and genes, respectively, the groups automatically introduce a block grouping structure on $\boldsymbol{B}_X$, with row blocks corresponding to gene groups and column blocks corresponding to ROI groups. They also induce the same gene grouping structures on $\boldsymbol{X} \times \mathbf{I}_{ad}$ and $\boldsymbol{X} \times \mathbf{I}_{mci}$, and the same block grouping structure on $\boldsymbol{B}_{Xad}$ and $\boldsymbol{B}_{Xmci}$. We assume that association signals are sparse at both group and individual levels. That is (i) each response group only associates with at most a few predictor groups, and (ii) each important voxel only associates with a few number of SNPs (SNP-disease interactions) compared to the sample size. In the following analyses, we assume that variables $\mathbf{I}_{ad}$, $\mathbf{I}_{mci}$, $\mathbf{Age}$ and $\mathbf{Sex}$ in model (2) each forms a group by itself.

2.1 First stage: selecting important ROI-to-gene blocks

In the first stage, we use the multivariate group lasso [29, 61] to select the important ROI-to-gene pairs. It serves as a screening step, which rules out the unimportant ROI-to-gene pairs by shrinking the corresponding association blocks to zero. To reduce the dimensionality of input variables while keeping the ROI and gene grouping structures, we use the major principle components (PC) within each ROI or gene group instead of using the voxel intensities and SNP genotypes. Note that PCs are linear combinations of the original variables, therefore a zero association block between the original variables implies a zero block between corresponding PCs. We interpret the selected PC association blocks as the evidence of associations between their representative ROIs and genes. The advantage of using PCs is two-fold. Firstly, it helps reduce the input dimensionality while keep the grouping structure and essential information within each group, therefore improves the efficiency of group-level selection. Secondly, since PCs are orthogonal (independent) to each other, they avoid the complications arising from collinearity between predictors or from overlapping grouping structures (genes could be overlapping with each other).

Let $\mathscr{R} = \{1, \ldots, R\}$ be the index set of ROI groups, and $\mathscr{G} = \{1, \ldots, G\}$ be the index set of generic predictor groups – i.e., gene, disease indicator, gene-disease interaction and other covariate groups. For ease of notation, we simply term each generic predictor group as a “gene group” in the following when no confusion is caused. Denote by $\mathscr{R} \otimes \mathscr{G}$ the induced block grouping structure on the regression coefficient matrix. For each $r \in \mathscr{R}$, denote by $\mathbf{P}_Y^r$ the major PCs of the responses in the rth group. Let $\mathbb{P}_Y = \left(\mathbf{P}_Y^1, \ldots, \mathbf{P}_Y^R\right)$ be the new response matrix of PCs. Similarly, for each $g \in \mathscr{G}$, denote by $\mathbf{P}_X^g$ the major PCs of the predictors in the gth group. Let $\mathbb{P}_X = \left(\mathbf{P}_X^1, \ldots, \mathbf{P}_X^G\right)$ be the new predictor matrix of PCs. We apply the multivariate group lasso on the PC matrices to select important ROI-to-gene associations by solving the optimization problem:

$$\underset{\boldsymbol{\Gamma}}{\operatorname{argmin}} \frac{1}{2n}\|\mathbb{P}_Y - \mathbb{P}_X\boldsymbol{\Gamma}\|_2^2 + \lambda_1 \sum_{rg \in \mathscr{R} \otimes \mathscr{G}} \omega_{rg}^{1/2}\|\boldsymbol{\Gamma}_{rg}\|_2, \qquad (3)$$

where $\|\cdot\|_2$ denotes the l_2 norm. Here $\boldsymbol{\Gamma}$ is the regression coefficient matrix between the PC matrices and $\boldsymbol{\Gamma}_{rg}$ is a submatrix block between rth ROI and gth gene group. The group lasso penalty $\sum_{rg \in \mathscr{R} \otimes \mathscr{G}} \omega_{rg}^{1/2}\|\boldsymbol{\Gamma}_{rg}\|_2$ aims to shrink the unimportant $\boldsymbol{\Gamma}_{rg}$ blocks to zero and ω_{rg} is a non-negative weight assigned to $\boldsymbol{\Gamma}_{rg}$, $r = 1, \ldots, R$, $g = 1, \ldots, G$. In our brain-wide GWAS, we use $\omega_{rg} = \sqrt{v \times s}$ [61, 44], where v is the number of voxels in the rth ROI and s is the number of predictors in the gth gene group. We set $\omega_{rg} = 0$ if we do not want to penalize the rgth group. The tuning parameter λ_1 controls the sparsity of the selected ROI-to-gene blocks.

2.2 Second stage: selecting important voxel-to-SNP signals

For each nonzero $\boldsymbol{\Gamma}_{rg}$ selected from the first stage, the corresponding ROI-to-gene pairs are passed to the second stage. In the second stage, we further zoom in to look at the associations for those pairs at voxle-to-SNP levels. For each selected ROI-to-gene pair, we solve the following multivariate lasso problem [29, 24, 18],

$$\underset{\boldsymbol{B}_{rg}}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{Y}_r - \boldsymbol{X}_g\boldsymbol{B}_{rg}\|_2^2 + \lambda_2 \sum_{\beta_{jk} \in \boldsymbol{B}_{rg}} \omega_{jk}|\beta_{jk}|, \qquad (4)$$

where the response variables $\boldsymbol{Y}_r$ are voxel-level intensity scores in the selected rth ROI, the predictors $\boldsymbol{X}_g$ are SNP genotypes (or SNP-disease interactions) within the selected associated gth gene group and $\boldsymbol{B}_{rg}$ is the corresponding regression coefficient block. Here λ_2 is a tuning parameter controlling the within-group individual-level sparsity, and ω_{jk} is a pre-assigned non-negative weight to β_{jk}. If $\omega_{jk} = 0$, then β_{jk} will not be penalized.

2.3 Stability selection and control for false discoveries

Stability selection [35] is employed in both stages. We fit models (3) and (4), respectively, multiple times, say K times, on randomly resampled (bootstrapped or subsampled) datasets using pre-fixed tuning parameters. Then an important signal (either group-level or individual-level) is eventually selected if its selection frequency among the K times of variable selection is greater than certain specified threshold.

The advantages of stability selection are three-fold. Firstly, it can reduce the random variation in the data coming from sampling or measurement error. Secondly, it saves the computing cost in choosing tuning parameters λ_1 and λ_2. Instead of using cross-validation to select optimal tuning parameters, the stability selection suggests to use a fixed set of tuning parameter values on re-randomized datasets. As long as the proposed fixed tuning parameter values are in a reasonable range, i.e., they are neither too large that shrink almost everything to zeros nor too small that barely shrink anything, the variable selection results are quite stable. Figure 3 in the online Appendix illustrates that the top signals identified in the analysis of ADNI PET imaging and genetic data are robustly selected when using bootstrapped samples and different values of the tuning parameters. The stability selection can be easily implemented and run on multi-core computing clusters and therefore is much more computationally efficient. Thirdly, stability selection provides a quantitative way to govern the number of false discoveries, for which we will discuss in detail in Section 4.

2.4 Selection properties

We show that the proposed structured brain-GWAS method achieves certain oracle bounds for selection, which are the selection bounds one could obtain as if the true model were given [4].

First, we introduce some notation. Let $\mathscr{J}_1(\mathbb{B}) = \left\{jk : \left|\beta_{jk}\right| \neq 0\right\}$ be the index set of nonzero elements in $\mathbb{B}$, and let $\mathscr{J}_2(\mathbb{B}) = \left\{rg \in \mathscr{R} \otimes \mathscr{G}, \|\boldsymbol{B}_{rg}\|_2 \neq 0\right\}$ be the index set of nonzero groups. Define $M_1(\mathbb{B}) = \sum_{jk} I\left(\beta_{jk} \neq 0\right) = |\mathscr{J}_1(\mathbb{B})|$ and $M_2(\mathbb{B}) = \sum_{rg \in \mathscr{R} \otimes \mathscr{G}} I\left(\|\boldsymbol{B}_{rg}\|_2 \neq 0\right) = |\mathscr{J}_2(\mathbb{B})|$. Denote by q_r the number of voxels in the rth ROI group and denote by p_g the number of predictors in the gth gene group. We assume that the predictors have a common marginal variance σ^2.

Next, we provide assumptions for the results given in Theorem 1.

i. Group-level generalized sparse condition (gGSC): For any $\eta_1 \geq 0$, there exists a non-empty set $\mathscr{A} \subset \mathscr{R} \otimes \mathscr{G}$, such that $\sum_{rg \in \mathscr{A}} \|\boldsymbol{B}_{rg}\|_2 \leq \eta_1$.

ii. Sparse Riesz condition (SRC): There exist spectrum bounds $0 < c_* < c^* < \infty$, such that for any $\mathscr{A}_1 \subset \{1, \ldots, G\}$ with rank $q^* = |\mathscr{A}_1|$ and any nonzero vector $\boldsymbol{\nu} \in \mathscr{R}^{\sum_{g \in \mathscr{A}_1} p_g}$, let $\mathbb{X}_{\mathscr{A}_1} = \left(\boldsymbol{X}_g, g \in \mathscr{A}_1\right)$ be the submatrix of $\mathbb{X}$ with its group indices in $\mathscr{A}_1$, the following inequalities hold

$$c_* \leq \frac{\|\mathbb{X}_{\mathscr{A}_1} \nu\|_2^2}{n\|\nu\|_2^2} \leq c^* \tag{5}$$

iii. Individual-level restricted eigenvalue condition (iREC): For any $\boldsymbol{B}_{rg} \in \mathscr{J}_2(\mathbb{B})$, suppose that $\boldsymbol{B}_{rg} \in \mathscr{R}^{p_g \times q_r}$. Let $\mathscr{J} \subseteq \left\{jk : 1 \leq j \leq p_g, 1 \leq k \leq q_r\right\}$ be any index set that satisfies $|\mathscr{J}| \leq s$ for some $0 < s \leq p_g \times q_r$. Then for any nontrivial matrix $\Delta \in \mathscr{R}^{p_g \times q_r}$ that satisfies $\left|\Delta_{\mathscr{J}^c}\right|_1 \leq 3\left|\Delta_{\mathscr{J}}\right|_1$, we have the following:

$$\kappa = \min_{\mathscr{J}, \Delta \neq 0, g \in \mathscr{G}} \frac{\|\boldsymbol{X}_g \Delta\|_2}{n^{1/2}\|\Delta_{\mathscr{J}}\|_2} > 0.$$

Here $\Delta_{\mathscr{J}}$ is the projection of Δ on an index set $\mathscr{J}$, that is the matrix with the same elements of Δ on coordinates $\mathscr{J}$ and zeros on the complementary coordinates $\mathscr{J}^c$.

iv. Let $d^* = \max_{rg \in \mathscr{R} \otimes \mathscr{G}} \omega_{rg}$, $d_* = \min_{rg \in \mathscr{R} \otimes \mathscr{G}} \omega_{rg}$ for ω_{rg}s in (3). Define $d = d^*/d_*$. Define $\eta_2 = \max_{\mathscr{A} \subset \mathscr{R} \otimes \mathscr{G}} \| \sum_{rg \in \mathscr{A}} \boldsymbol{X}_g \boldsymbol{B}_{rg} \|_2$,

$r_1 = \left(\frac{nc^* \sqrt{d^*} \eta_1}{\lambda_1 d_* M_2} \right)^{1/2}$, $r_2 = \left(\frac{nc^* \eta_2^2}{\lambda_1^2 d_* M_2} \right)^{1/2}$, $\bar{c} = c^*/c_*$ and Let $\sigma * = \sigma \sqrt{\max_{g \in \mathscr{G}} p_g}$. Assume that λ_1 $C_2 = 2 + 4r_1^2 + 4\sqrt{d\bar{c}} r_2 + 4d\bar{c}$.

in model (3) satisfies

$$\lambda_1 \geq \max\left\{ \lambda_0, \lambda_{n,G} \right\},$$

where $\lambda_{n,G} = 2\sigma^* \sqrt{8(1 + c_0) d_* d^2 q^* \bar{c} n c^* \log(N_d \vee a_n)}$ with $N_d = \sum_{rg \in \in \mathscr{R} \otimes \mathscr{G}} \omega_{rg} c_0 \geq 0$ and $a_n \geq 0$ satisfying $d_* G/(N_d \vee a_n)^{1 + c_0} \approx 0$, and $\lambda_0 = \inf\{\lambda : C_2 M_2(\mathbb{B}) + 1 \leq q * \}$ with $\inf \varnothing = \infty$. Here $a \vee b = \max\{a, b\}$.

Theorem 1.—Theorem 1. *Let* $\mathbb{B} *$ *be the true coefficient matrix. Assume that each of the* $\mathbb{X}$ *variables has mean 0 and marginal variance* $\sigma^2 = 1$. *Let* $\psi_{\max}$ *be the largest eigenvalue of* $\mathbb{X}^{\mathrm{T}} \mathbb{X}/n$. *Denote* $M_1^*(\mathbb{B}^*) = \max_{rg \in \mathscr{R} \otimes \mathscr{G}} M_1\left(\boldsymbol{B}_{rg}^*\right)$. *Assume gGSC, SRC and iREC hold. Then with probability converging to 1 as* $n \to \infty$, *we have the following oracle selection bounds for group- and individual-level signals:*

$$M_2(\widehat{\mathbb{B}}) \leq C_2 M_2(\mathbb{B} *), \tag{6}$$

$$M_1(\widehat{\mathbb{B}}) \leq 64 \psi_{\max} C_2 M_2\left(\mathbb{B}^*\right) M_1^*\left(\mathbb{B}^*\right)/\kappa^2. \tag{7}$$

When gGSC, SRC and (iv) hold, Wei and Huang [55] showed that the group-level selection bound holds for the univariate-response group lasso. The proof of (6) follows the first assertion in Theorem 2.1 in [55], execpt that we need to show that SRC holds for $\mathbb{P}_X$, as in our method the group lasso is applied to $\mathbb{P}_X$ instead of $\mathbb{X}$ in the first stage. In fact, since each PC is a linear combination of the orignal $\mathbb{X}$ variables, we can write $\mathbb{P}_{X, \mathscr{A}_1} = \mathbb{X}_{\mathscr{A}_1} \mathbf{W}$, where $\mathbf{W}$ is a $P \times R$ weight matrix, $R \leq P$, consisting of the eigenvectors of the covariance matrix of $\mathbb{X}$. Then we have $\|\mathbb{P}_{X, \mathscr{A}_1} \boldsymbol{\nu}\|_2^2 / \left\{ n \|\nu\|_2^2 \right\} = \|\mathbb{X}_{\mathscr{A}_1} \mathbf{W} \nu\|_2^2 / \left\{ n \|\nu\|_2^2 \right\} = \|\mathbb{X}_{\mathscr{A}_1} \mathbf{W} \nu\|_2^2 / \left\{ n \nu^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \mathbf{W} \nu \right\} = \|\mathbb{X}_{\mathscr{A}_1} \nu'\|_2^2 / \{ n \|\nu'\|_2^2 \}$, where $\boldsymbol{\nu}' = \mathbf{W} \boldsymbol{\nu}$. Therefore the SCR holds for $\mathbb{P}_X$ if it holds for $\mathbb{X}$. The individual-level oracle selection bound (7) directly follows from (6) and the multivariate lasso oracle selection bound introduced in Theorem 2 in Li et al. [29].

3 A simulation study

We investigate the empirical selection performance for the proposed two-stage method through simulations. Assume that both $\boldsymbol{Y}$ and $\mathbb{X}$ have 50 groups with each group containing 200 variables.

The coefficient matrix $\mathbb{B}$ assumes a block diagonal structure, i.e., the 1st $\boldsymbol{Y}$ group is associated with only the 1st $\mathbb{X}$ group, the 2nd $\boldsymbol{Y}$ group is associated with only the 2nd $\mathbb{X}$ group, etc. Coefficients within off-diagonal blocks are set to be zeros. Half of the coefficients within diagonal blocks are randomly generated from Unif([−5,−3]U[3,5]) and the other half are set as zeros (therefore, the sparsity within important coefficient blocks is 0.5). Once $\mathbb{B}$ is generated, it is fixed in all experiments.

We assume $\mathbb{X}$ groups are uncorrelated. Within-group $\mathbb{X}$ variables are generated from a multivairate normal distribution with zero means and a first-order auto-correlation structure with a correlation coefficient 0.5, denoted by AR1(0.5), and unit marginal variances.

We generate the noise variables **E** from a multivairate normal distribution with the following three correlation structures and unit marginal variances.

i. Independent $\boldsymbol{Y}$ groups: Variables within each $\boldsymbol{Y}$ group take an AR1(0.5) correlation structure.

ii. Weakly correlated $\boldsymbol{Y}$ groups: Variables within each $\boldsymbol{Y}$ group take an AR1(0.5) correlation structure. Variables from different $\boldsymbol{Y}$ groups are correlated with a compound symmetry (CS) correlation structure with a coefficient 0.1, denoted by CS(0.1). Therefore, the overall $\boldsymbol{Y}$ correlation structure is CS(0.1)⊗AR1(0.5), where ⊗ is the Kronecker product.

iii. Moderately correlated $\boldsymbol{Y}$ groups: Variables within each $\boldsymbol{Y}$ group take an AR1(0.5) correlation structure. Variables from different $\boldsymbol{Y}$ groups take a CS(0.5) correlation structure. The overall $\boldsymbol{Y}$ correlation structure is CS(0.5)⊗AR1(0.5).

The response matrix is then generated according to $Y = \mathbb{X}\mathbb{B} + \mathbf{E}$. For each scenario, we generate datasets with three different sample sizes n = 200, 500 and 1000.

For each simulated dataset, the proposed method is applied in each stage followed by stability selections. In the first stage, we use major PCs in each response/predictor group that explain more than 80% of the total within-group variation. Each stability selection is carried out on 100 bootstrapped datasets. Optimal tuning parameters are selected by five-fold cross-validation for each stage of selection. Tuning parameters are then fixed in the stability selection. Selection frequency threshold is set to be 80% for both stages. One hundred independent experiments are repeated for each setting. We report means and empirical standard deviations for Sensitivity (SE) and Specificity (SP) in Table 1. The first stage group-level SE and SP are defined by:

$$\mathrm{SE}(1) = \frac{|\{rg : 1 \le r \le R, 1 \le g \le G, \left\|\widehat{\boldsymbol{\Gamma}}_{rg}\right\|_2 \ne 0 \text{ and } \|\boldsymbol{B}^*_{rg}\|_2 \ne 0\}|}{|\{rg : 1 \le r \le R, 1 \le g \le G, \|\boldsymbol{B}^*_{rg}\|_2 \ne 0\}|} \text{ and}$$

$$\mathrm{SP}(1) = \frac{|\{rg : 1 \le r \le R, 1 \le g \le G, \|\widehat{\boldsymbol{\Gamma}}_{rg}\|_2 = 0 \text{ and } \|\boldsymbol{B}^*_{rg}\|_2 = 0\}|}{|\{rg : 1 \le r \le R, 1 \le g \le G, \|\boldsymbol{B}^*_{rg}\|_2 = 0\}|},$$

where the superscript $*$ representing the true values. And the second stage individual-level SE and SP are defined by:

$$\mathrm{SE}(2) = \frac{|\{jk : 1 \le j \le P, 1 \le k \le Q, \widehat{\beta}_{jk} \ne 0 \text{ and } \beta^*_{jk} \ne 0\}|}{|\{jk : 1 \le j \le P, 1 \le k \le Q, \beta^*_{jk} \ne 0\}|} \text{ and}$$

$$\mathrm{SP}(2) = \frac{|\{jk : 1 \le j \le P, 1 \le k \le Q, \widehat{\beta}_{jk} = 0 \text{ and } \beta^*_{jk} = 0\}|}{|\{jk : 1 \le j \le P, 1 \le k \le Q, \beta^*_{jk} = 0\}|}.$$

For comparison, we also conduct pairwise marginal linear regressions followed by Bonferroni correction for multiple comparisons. The $\widehat{\beta}_{jk}$s with p-values less than the Bonferroni corrected threshold (5e-12) are selected as important signals. The results for the pairwise approach are given in the last two columns in Table 1.

The simulation results show that our two-stage method combined with stability selection renders very good selection results for group structured ultrahigh-dimensional multivariate responses and multiple predictors data. It is far more powerful than the pairwise approach. Especially for the first-stage group-level selection, our approach gives almost perfect selection performance even when the sample size is very small. For the second-stage individual-level selection, the selection performance improves significantly as the sample size increases. The selection performance is similar in all three different responses' correlation structures.

4 Analysis of ADNI FDG-PET and SNP data

The ADNI data used in our structured brain-GWAS analysis contains three parts: imaging data, genetic data and clinical data, all from the ADNI database. Samples with both imaging and genotype data are included in the analysis, resulting in a dataset with 373 samples including 86 AD patients, 188 MCI patients and 99 normal controls (NC). The clinical data contain the disease status (AD, MCI or NC), demographic information (e.g. age and sex) and $\epsilon 4$ allele information for the apolipoprotein E (*APOE*) gene. We fit model (1) to the ADNI PET imaging and genomic data using the proposed method.

4.1 PET images and ROI's

Images used in our analysis are FDG-PET images, which have been widely used in neuroimaging studies for over 20 years. FDG-PET images measure cerebral glucose metabolic activities. From year 2003 to 2011, a total of 403 FDG-PET scans have been acquired at approximately 50 different participating sites in ADNI-1 and ADNI-GO studies, including 95 AD subjects, 206 MCI subjects and 102 NC subjects. Due to missing genetic information, only 373 individuals are included in our study. Each image contains 349,182 voxels embedded in a 160×160×96 3D array. Those images were preprocessed to produce a uniform isotropic resolution.

To incorporate the brain anatomic structures, the PET images were segmented by Brodmann atlas [8]. As a result, voxels in each image were grouped into 106 Brodmann ROIs. Voxels not indexed by the Brodmann atlas are not considered in the analysis. The regions on the left hemisphere are symmetric mirror reflection of the ones on the right hemisphere. In the following, we use "(L)" to denote the regions on the left hemisphere and "(R)" to denote the regions on the right hemisphere. For example, "Temporal cortex BA20(L)" refers to the temporal cortex region named "BA20" on the left hemisphere and "Temporal cortex_BA20(R)" refers to the corresponding symmetric region on the right hemisphere.

4.2 Genotypes

ADNI SNP data were genotyped using Illumina 610 Quad array with more than 620,000 tag SNPs. Genotyping was performed by Polymorphic DNA Technologies. We grouped SNP genotypes into genes using the UCSC known genes list of NCBI36 assembly (http://genome.ucsc.edu), with each gene containing the SNPs within its physical range plus a flanking region of 100 KB up- and down-streams. This resulted in a total of 29,458 genes in the 22 autosomes. For isoform genes, we took the joint regions of all the isoforms to be the same gene.

The raw genotypes were screened by a series of quality control procedures. SNPs with missing rates greater than 1%, heterozygous haploid and markers with Hardy-Weinberg equilibrium p-values less than 10^{-6} were removed, which left in a total of 564,636 SNPs in the analysis. The missing genotypes with missing rate under 1% were imputed by the average genotype scores of the non-missing genotypes.

4.3 Data analysis

In the first-stage selection, we use the first five PCs in each brain ROI and the first twenty PCs or the first several PCs that explain at least 80% variation, whichever smaller, in each gene. Most of the ROIs have more than 70% of their variations explained by their first five PCs. Most of the genes have at least 80% of variations explained no more than 20 PCs. For example, only seven out of 800 genes on chromosome 20 have less than 60% of variations explained by their first 20 PCs. Figure 1 in the online Appendix shows the percentage of total variation explained by the first five PCs in each ROI and the percentage of variation explained by up to the first 20 PCs in each gene on chromosome 20. The $\epsilon 4$ allele of the apolipoprotein E gene ($APOE$-$\epsilon 4$) is the most common genetic risk factor for AD [13, 48]. However, ADNI genetic dataset does not contain the genotypes for the SNPs in the $APOE$-$\epsilon 4$ gene. We extract the $APOE$-$\epsilon 4$ allele information score from ADNI clinical data and put it together with the first 20 PCs on chromosome 19.

We use the R package MSGLasso [30] to run the multivariate group lasso on the PC matrices. Stability selection [35] is then performed on 100 bootstrapped datasets. ROI-to-gene pairs with at least 75% stability selection frequency are selected as important ROIs and genes in the first selection stage. For the *APOE* gene, we use *APOE-ϵ4* allele score to fit model (4), wherever *APOE* is selected.

Meinshausen and Bühlmann [35] showed that the expected number V of falsely selected variables is bounded from above by

$$E(V) \leq \frac{1}{2\pi_{thr} - 1}\frac{q^2}{P}, \tag{8}$$

where π_{thr} is the thresholding frequency used for the selection, which in our case is 75% for the first stage and 80% for the second stage, and q is the average number of selected variables. In our study, the typical numbers of selected variables are from tens to hundreds out of tens of thousands of variables in total, which yield $q^2/P \ll 1$. Therefore the error number per chromosome is controlled by $\ll 1/(2 \times 0.75 - 1) = 2$. That is, for each ROI, in the first stage, there will be just a few falsely discovered genes across all chromosomes.

In the second stage, we also use MSGLasso [30] to fit a multivariate lasso regression on each of the selected ROI-to-gene pairs. Stability selection is then carried out on 100 bootstrapped datasets for each ROI-to-gene pair. Voxel-to-SNP pairs with selection frequency greater than 80% are selected to be important individual-level signals. Then we apply a multiple linear regression for each selected voxel with its selected important SNP predictors for post-selection estimation and inference. In our ADNI data analysis, the typical number of important SNPs selected for a voxel ranges from a few to several dozens, which is much smaller than the sample size.

In both stages, we do not penalize on $\mathbf{I}_{ad}$, $\mathbf{I}_{mci}$, **Age** and **Sex** by setting the corresponding $\omega_{gr} = 0$ or $\omega_{jk} = 0$.

4.4 Results

Table 2 provides a list of top signals that meet both criteria of p-values less than 10^{-6} and selection frequencies greater than 80%. The selected brain regions and strength of the SNP effects are also illustrated in Figure 2. Since there is no SNP-MCI interaction effect satisfying both criteria, we provide a list of top MCI interactions in Table 1 in the online Appendix. Table 2 in the online Appendix lists the top selected ROIs and voxels therein for the *APOE-ϵ4* effects.

Some brain regions are identified to have either significant gene effects or gene-AD interaction effects. For example, regions, such as BA40(L), BA39(R), BA39(L), BA7(R) and BA7(L) in the superior parietal cortices are found significantly associated with certain genes or with their associations significantly modified by the AD status. On the contrary, no genome-wide significant SNP was found in the previous pair-wise brain-GWASs [46]. We have confirmed some brain regions associated with genetics appeared in the existing literature. For example, Mills et al. [36] reported associations between lipid metabolism in superior parietal cortices and alternatively spliced isoforms in RNA transcriptome. Other identified regions that are associated with genetics or with their genetic effects significantly modified by AD status include BA18(R), BA18(L), BA19(R), BA19(L) in occipital cortices

[7] and BA20(R), BA20(L), BA21(R), BA21(L), BA22(R) and BA22(L) in temporal cortices [47, 42, 7].

Some genetic findings in previous studies are confirmed by our brain-GWAS. For example, Wang et al. [54] found that inhibiting *IL8RB* (*CXCR2*) can turn down anyloid-β production and protect neural cells. Nakamura et al. [38] found a similar effect of *COLEC12* (*SRCL*) gene in AD samples. Other direct supports on AD interactions include Burns et al. [10] on *SAKCA* (*KCNMA1*), Xie et al. [58] on *PRIMA*, Nakamura et al. [38] on *COLEC12* and Broer et al. [9] on *HSPA13*.

Some gene-to-AD interactions have also been found in the literature to be associated with other cognitive-related diseases such as autism and hearing impairment. Such genes include *AK096399* [11], *GJB2* [32], *SNX29* [50], *MED1* [20, 56] and *COL9A3* [45, 1].

We also confirm some gene effects on brain metabolizing. For example, *CDC42EP3* encodes certain family of guanosine triphosphate metabolizing proteins and the gene is weakly expressed in brain. *PACS2* plays a role in membrane traffic with tumour-necrosis-factor-related apoptosisinducing-ligand (TRAIL) induced apophasis [2], which in turn can cause human brain cell death [39].

Our findings also provide evidence about indirect genetic effects on certain chemical compound or protein translocation, which are reflected in the PET scans and may be associated with AD. For example, Dai et al. [15] and Sakamoto and Holman [43] demonstrate that *TBC1D4* plays a role in regulation of GluT4 traffic, which, on the other hand is associated with AD [49, 60]. Nolte et al. [40] and Lu et al. [33] give a chain of relationships of *HOXD4* gene to Pax6 protein to AD.

There are also several novel signals which have not been reported in previous literature, such as associations between *BC007399* and BA39(R) in the superior parietal cortex, between *GALNT4* and BA19(L) in the occipital cortex and between *RIN2* and CERHEM(L).

5 Discussion

The overall computational cost of our two-stage approach is lower than the pairwise approaches [46, 19], as our approach removes the unimportant ROI-to-gene signal blocks first and only focuses on the selected ROI-to-gene blocks in the downstream analysis stages. To further save the computational time, we parallelized the computational jobs on multi-core computing clusters. Our approach has more power also due to the integration of the brain and genome grouping structures. In Stein et al. [46], no significant voxel-to-SNP signals were found due to the huge number of multiple comparisons.

We recognize that post-selection inference is biased. Simultaneous selection, estimation and inference have been studied recently [53, 3]. Kuchibhotla et al. [26] also provide an upper bound for post-selection inference p-values when taking into account the selection bias. These will be investigated for our proposed method in future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the two reviewers and Dr. Linglong Kong for helpful suggestions. Nan's research was supported in part by NIH R01 AG056764 and NSF DMS 1915711. Zhu's research was supported in part by NSF DMS 1821243.

References

[1]. Asamura K, Abe S, Fukuoka H, Nakamura Y, and Usami S (2005). Mutation analysis of COL9A3, a gene highly expressed in the cochlea, in hearing loss patients. Auris Nasus Larynx., 32(2), 113–117. [PubMed: 15917166]

[2]. Aslan J, You H, Williamson DM, Endig J, amd L Thomas RY, Shu H, Du Y, Milewski R, Brush M, Possemato A, Sprott K, Fu H, Greis K, Runckel D, Vogel A, and Thomas G (2009). Akt and 14–3-3 control a *PACS-2* homeostatic switch that integrates membrane traffic with trail-induced apoptosis. Mol Cell, 34(4), 497–509. [PubMed: 19481529]

[3]. Berk R, Brown L, Buja A, Zhang K, and Zhao L (2013). Valid post-selection inference. Ann. Statist, 41(2), 802–837.

[4]. Bickel PJ, Ritov Y, and Tsybakov AB (2009). Simultaneous analysis of Lasso and Danzig selector. Ann. Statist, 37, 1705–1732.

[5]. Biffi A, Anderson C, Desikan R, Sabuncu M, Cortellini L, Schmansky N, Salat D, Rosand J, and ADNI (2010). Genetic variation and neuroimaging measures in Alzheimer disease. Arch Neurol, 67(6), 677–685. [PubMed: 20558387]

[6]. Braber A, Bohlken M, Brouwer R, Ent D, Kanai R, Kahn R, Geus E, Pol H, and Boomsm D (2013). Heritability of subcortical brain measures: A perspective for future genome-wide association studies. NeuroImage, 83, 98–102. [PubMed: 23770413]

[7]. Braskie N, Ringman J, and Thompson P (2011). Neuroimaging measures as endophenotypes in Alzheimer's disease. International Journal of Alzheimer's Disease. 10.4061/2011/490140.

[8]. Brodmann K (2010). Brodmann's Localisation in the Cerebral Cortex. Springer.

[9]. Broer L, Ikram M, Schuur M, DeStefano A, Bis J, Liu F, Rivadeneira F, Uitterlinden A, Beiser A, Longstreth W, Hofman A, Aulchenko Y, Seshadri S, Fitzpatrick A, Oostra B, Breteler M, and van Duijn C (2011). Association of *HSP70* and its co-chaperones with Alzheimer's disease. J Alzheimers Dis, 25(1), 93–102. [PubMed: 21403392]

[10]. Burns L, Minster R, Demirci F, Barmada M, Ganguli M, Lopez OL, DeKosky S, and Kamboha M (2011). Replication study of genome-wide associated SNPs with late-onset Alzheimer's disease. Am J Med Genet B Neuropsychiatr Genet, 156(4), 507–512.

[11]. Cannon D, Miller J, Robison R, Villalobos M, Wahmhoff N, Allen-Brady K, McMahon W, and Coon H (2010). Genome-wide linkage analyses of two repetitive behavior phenotypes in utah pedigrees with autism spectrum disorders. Molecular Autism, 1(1), 3. doi: 10.1186/2040-2392-1-3. [PubMed: 20678246]

[12]. Chilumuri A and Milton N (2013). The role of neurotransmitters in protection against amyloid-β toxicity by kiss-1 overexpression in SH-SY5Y neurons. Neuroscience, 10.1155/2013/253210.

[13]. Corder EH, Ghebremedhin E, Taylor MG, Thal DR, Ohm TG, and Braak H (2004). The biphasic relationship between regional brain senile plaque and neurofibrillary tangle distributions: modification by age, sex, and APOE polymorphism. Ann N Y Acad Sci, 1019, 24–28. [PubMed: 15246987]

[14]. Chung P, Beyens G, Boonen S, Papapoulos S, Geusens P, Karperien M, Vanhoenacker F, Verbruggen L, Fransen E, Offel JV, Goemaere S, Zmierczak H, Westhovens R, Devogelaer J, and Hul WV (2010). The majority of the genetic risk for paget's disease of bone is explained by genetic variants close to the *CSF1*, *OPTN*, *TM7SF4*, and TNFRSF11A genes. Human genetics, 128, 615–26. [PubMed: 20839008]

[15]. Dai M, Freeman B, Shikani HJ, Bruno FP, Collado JE, Macias R, Reznik SE, Davies P, Spray DC, Tanowitz HB, Weiss LM, and Desruisseaux MS (2013). Altered regulation of akt signaling with murine cerebral malaria, effects on long-term neuro-cognitive function, restoration with lithium treatment. PLoS ONE, 7(10), e44117. doi: 10.1371/journal.pone.0044117.

[16]. Desbaillets I, Diserens A, Tribolet N, Hamou M, and Meir EV (1997). Upregulation of interleukin 8 by oxygen-deprived cells in glioblastoma suggests a role in leukocyte activation, chemotaxis, and angiogenesis. J Exp Med, 186(8), 1201–1212. [PubMed: 9334359]

[17]. Ertekin-Taner N (2010). Genetics of Alzheimer disease in the pre- and post-GWAS era. Alzheimer's Research & Therapy, 2(1), 3. doi: 10.1186/alzrt26.

[18]. Friedman J, Hastie T and Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. [PubMed: 20808728]

[19]. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. Neuroimage, 63(2):858–873. [PubMed: 22800732]

[20]. Giordano A and Macaluso M (2011). Cancer Epigenetics: Biomolecular Therapeutics in Human Cancer. John Wiley & Sons.

[21]. Heikaus S, Winterhager E, Traub O, and Grümmer R (2002). Responsiveness of endometrial genes *Connexin26*, *Connexin43*, *C3* and clusterin to primary estrogen, selective estrogen receptor modulators, phyto- and xenoestrogens. J Mol Endocrinol, 29(2), 239–249. [PubMed: 12370124]

[22]. Hibar D, Stein J, Kohannim O, Jahanshad N, Saykin A, Shen L, Kim S, Pankratz N, Foroud T, Huentelman M, Potkin S, Jack CJ, Weiner M, Toga A, Thompson P, and ADNI (2011). Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. Neuroimage, 56(4), 1875–1891. [PubMed: 21497199]

[23]. Horuk R, Martin A, Wang Z, Schweitzer L, Gerassimides A, Guo H, Lu Z, Hesselgesser J, Perez H, Kim J, Parker J, Hadley T, and Peiper S (1997). Expression of chemokine receptors by subsets of neurons in the central nervous system. J Immunol. 1997, 158(6), 2882–2890.

[24]. Kohannim O, Hibar D, Stein J, Jahanshad N, Hua X, Rajagopalan P, Toga A, Jack CJ, Weiner M, de Zubicaray G, McMahon K, Hansell N, Martin N, Wright M, Thompson P, and ADNI (2012). Discovery and replication of gene influences on brain structure using lasso regression. Front Neurosci, 6, 115. [PubMed: 22888310]

[25]. Korac J, Schaeffer V, Kovacevic I, Clement A, Jungblut B, Behl C, Terzic J, and Dikic I (2013). Ubiquitin-independent function of optineurin in autophagic clearance of protein aggregates. J Cell Sci, 126(02), 580–92. [PubMed: 23178947]

[26]. Kuchibhotla AK, Brown LD, Buja A, George EI, and Zhao L (2013). Valid Post-selection Inference in Assumption-lean Linear Regression. arxiv.org, https://arxiv.org/pdf/1806.04119.pdf.

[27]. Kukull W, Schellenberg G, Bowen J, McCormick W, Yu C, Teri L, Thompson J, O'Meara E, and Larson E (1996). Apolipoprotein e in Alzheimer's disease risk and case detection: a case-control study. J Clin Epidemiol, 49(10), 1143–1148. [PubMed: 8826994]

[28]. Li Y, Hong HG, and Li Y (2019). Multiclass linear discriminant analysis with ultrahigh-dimensional features. Biometrics, 75(4), 1086–1097. DOI: 10.1111/biom.13065. [PubMed: 31009070]

[29]. Li Y, Nan B, and Zhu J (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics, 71(2), 354–363. DOI: 10.1111/biom.12292. [PubMed: 25732839]

[30]. Li Y, Nan B, and Zhu J (2016). MSGLasso: Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure. CRAN, https://cran.r-project.org/web/packages/MSGLasso/index.html.

[31]. Liu Y, Guo D, Tian L, Shang D, Zhao W, Li B, Fang W, Zhu L, and Chen Y (2010). Peripheral t cells derived from Alzheimer's disease patients overexpress *CXCR2* contributing to its transendothelial migration, which is microglial TNF-alpha-dependent. Neurobiol Aging, 31(2), 175–188. [PubMed: 18462836]

[32]. Lingala HB, Sankarathi, and Penagaluru PR (2009). Role of connexin 26 (GJB2) & mitochondrial small ribosomal RNA (mt 12S rRNA) genes in sporadic & aminoglycoside-

induced non syndromic hearing impairment. Indian J Med Res, 130, 369–378. [PubMed: 19942739]

[33]. Lu Y, He X, and Zhong S (2007). Cross-species microarray analysis with the OSCAR system suggests an *INSR→Pax6→NQO1* neuro-protective pathway in aging and Alzheimer's disease. Nucleic Acids Res, 35, W105–114. [PubMed: 17545194]

[34]. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease Alzheimer's & Dementia, 7(3), 263–269.

[35]. Meinshausen N and Bühlmann P (2010). Stability selection. J. R. Statist. Soc. B, 72, 417–473.

[36]. Mills J, Nalpathamkalam T, Jacobs H, Merico CJD, Hu P, and Janitz M (2013). RNA-seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. Neurosci Lett, 536, 90–95. [PubMed: 23305720]

[37]. Mosconi L (2005). Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. European Journal of Nuclear Medicine and Molecular Imaging, 32, 466–510.

[38]. Nakamura K, Ohya W, Funakoshi H, Sakaguchi G, Kato A, Takeda M, Kudo T, and Nakamura T (2006). Possible role of scavenger receptor *SRCL* in the clearance of amyloid-beta in Alzheimer's disease. J Neurosci Res, 84(4), 874–490. [PubMed: 16868960]

[39]. Nitsch R, Bechmann I, Deisz R, Haas D, Lehmann T, Wendling U, and Zipp F (2000). Human brain-cell death induced by tumour-necrosis-factor-related apoptosis-inducing ligand (trail). Lancet, 356(9232), 827–828. [PubMed: 11022932]

[40]. Nolte C, Rastegar M, Amores A, Bouchard M, Grote D, Maas R, Kovacs E, Postlethwait J, Rambaldi I, Rowan S, Yan Y, Zhang F, and Featherstone M (2006). Stereospecificity and *PAX6* function direct *Hoxd4* neural enhancer activity along the anteroposterior axis developmental biology. Developmental Biology, 299(2), 582–593. [PubMed: 17010333]

[41]. Peper J, Brouwer R, Boomsma D, Kahn R, and Pol H (2007). Genetic influences on human brain structure: A review of brain imaging studies in twins. Human Brain Mapping, 28(6), 464–473. [PubMed: 17415783]

[42]. Risacher S, West ASJ, Shen L, Firpi H, and McDonald B (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr Alzheimer Res, 6(4), 347–361. [PubMed: 19689234]

[43]. Sakamoto K and Holman GD (2008). Emerging role for *AS160*/*TBC1D4* and *TBC1D1* in the regulation of *GLUT4* traffic. Am J Physiol Endocrinol Metab, 295, e29–37. [PubMed: 18477703]

[44]. Silver M, Montana G, and Alzheimer's-Disease-Neuroimaging-Initiative (2012). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat Appl Genet Mol Biol, 11(1), 1–43.

[45]. Solovieva S, Lohiniva J, Leino-Arjas P, Raininko R, Luoma K, Ala-Kokko L, and Riihimäki H (2006). Intervertebral disc degeneration in relation to the *COL9A3* and the *IL-1ss* gene polymorphisms. Eur Spine J, 15(5), 613–619. [PubMed: 16133074]

[46]. Stein J, Hua X, Lee S, Ho A, Leow A, Toga A, Saykin A, Shen L, Foroud T, Pankratz N, Huentelman M, Craig D, Gerber J, Allen A, Corneveaux J, Dechairo B, Potkin S, Weiner M, Thompson P, and Initiative ADN (2010). Voxelwise genome-wide association study (vgwas). Neuroimage, 53(3), 1160–1174. [PubMed: 20171287]

[47]. Stein J, Hua X, Morra J, Lee S, Hibar D, Ho A, Leow A, Toga A, Sul J, Kang H, Eskin E, AJ AS, Shen L, Foroud T, Pankratz N, Huentelman M, Craig D, Gerber J, Allen A, Corneveaux J, DA DS, Webster J, DeChairo B, Potkin S, Jack C, Weiner M, Thompson P, and ANDI (2010b). Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. Neuroimage, 51(2), 542–554. [PubMed: 20197096]

[48]. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, and et al. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Neuroimage, 90, 1977–1981.

[49]. Talbot K, Wang H, Kazi H, Han L, Bakshi KP, Stucky A, Fuino RL, Kawaguchi KR, Samoyedny AJ, Wilson RS, Arvanitakis Z, Schneider JA, Wolf BA, Bennett DA, Trojanowski JQ, and

Arnold1 SE (2012). Demonstrated brain insulin resistance in Alzheimer's disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline. Journal of Clinical Investigation, 122(4), 1316–1338.

[50]. Teasdale RD and Collins BM (2012). Insights into the PX (phox-homology) domain and SNX (sorting nexin) protein families: structures, functions and roles in disease. Biochem. J, 441, 39–59. [PubMed: 22168438]

[51]. Tsai H, Frost E, To V, Ffrench-Constant SRC, Geertman R, Ransohoff R, and Miller R (2002). The chemokine receptor *CXCR2* controls positioning of oligodendrocyte precursors in developing spinal cord by arresting their migration. Cell, 110(3), 373–383. [PubMed: 12176324]

[52]. Vallès A, Grijpink-Ongering L, de Bree F, Tuinstra T, and Ronken E (2006). Differential regulation of the *CXCR2* chemokine network in rat brain trauma: implications for neuroimmune interactions and neuronal survival. Neurobiol Dis, 22(2), 312–322. [PubMed: 16472549]

[53]. van de Geer S, Bühlmann P, Ritov Y and Dezeure R (2014). On asymptotically optinal confidence regions and tests for high-dimensional models. Ann. Statist, 42(3), 1166–1202.

[54]. Wang J, Shi Z, Xu X, Xin G, Chen J, Qi L, and Li P (2013). Triptolide inhibits amyloid-β production and protects neural cells by inhibiting *CXCR2* activity. J Alzheimers Dis, 33(1), 217–229. [PubMed: 22986777]

[55]. Wei FR and Huang J (2015). Consistent group selection in high-dimensional linear regression. Bernoulli, 16(4), 1369–1384. DOI: 10.3150/10-BEJ252.

[56]. Wong CCY, Meaburn EL, Ronald A, Price TS, Jeffries AR, Schalkwyk LC, Plomin R, and Mill J (2013). Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. Molecular Psychiatry, doi: 10.1038/mp.2013.41.

[57]. Xia M, Qin S, McNamara M, and Hyman BT (1996). Type b IL8 receptor (IL8RB) in neuritic plaques of Alzheimer's disease. Journal of Neuropathology and Experimental Neurology, 55(5).

[58]. Xie H, D DL, Leung K, Chen V, Zhu K, Chan W, Choi R, Massoulié J, and Tsim K (2010). Targeting acetylcholinesterase to membrane rafts: a function mediated by the proline-rich membrane anchor (PRiMA) in neurons. J Biol Chem, 285(15), 11537–11546. [PubMed: 20147288]

[59]. Yaffe K, Krueger K, Cummings SR, Blackwell T, Henderson VW, Sarkar S, Ensrud K, and Grady D (2005). Effect of raloxifene on prevention of dementia and cognitive impairment in older women: The multiple outcomes of raloxifene evaluation (MORE) randomized trial. Am J Psychiatry, 162, 683–690. [PubMed: 15800139]

[60]. Yang J, Li S, and Liu Y (2013). Systematic analysis of diabetes- and glucose metabolism-related proteins and its application to Alzheimer's disease. J. Biomedical Science and Engineering, 6, 615–644.

[61]. Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B, 68, 49–67.

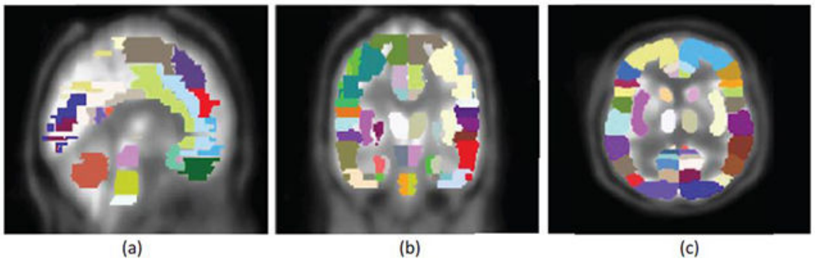
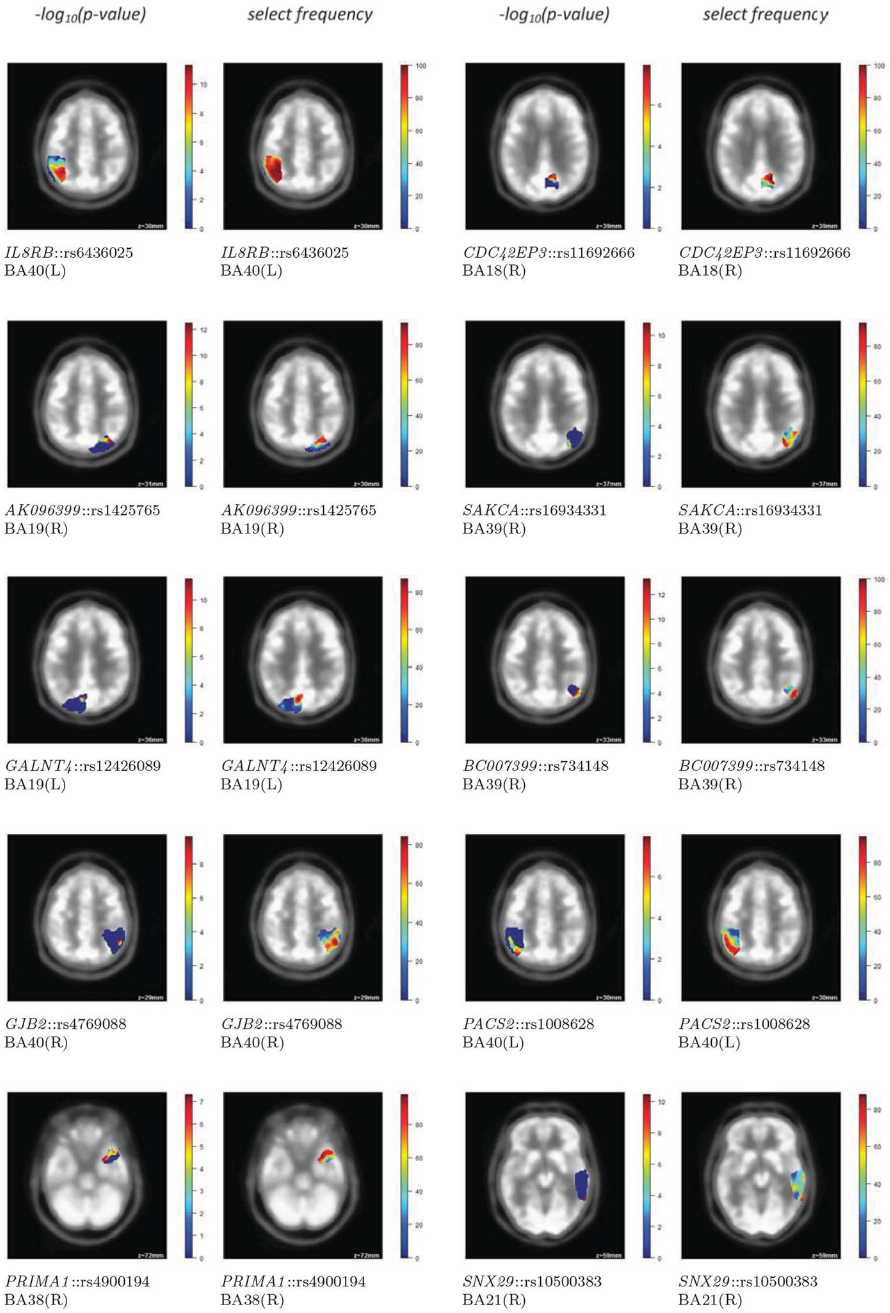


Figure 1:

Illustration of mapping Brodmann atlas of ROIs onto segmented PET images. ROIs are highlighted with colors. (a) Sagittal slice at midline. (b) Coronal slice at midline. (c) Axial slice at midline.



IL8RB::rs6436025 BA40(L)

IL8RB::rs6436025 BA40(L)

CDC42EP3::rs11692666 BA18(R)

CDC42EP3::rs11692666 BA18(R)

AK096399::rs1425765 BA19(R)

AK096399::rs1425765 BA19(R)

SAKCA::rs16934331 BA39(R)

SAKCA::rs16934331 BA39(R)

GALNT4::rs12426089 BA19(L)

GALNT4::rs12426089 BA19(L)

BC007399::rs734148 BA39(R)

BC007399::rs734148 BA39(R)

GJB2::rs4769088 BA40(R)

GJB2::rs4769088 BA40(R)

PACS2::rs1008628 BA40(L)

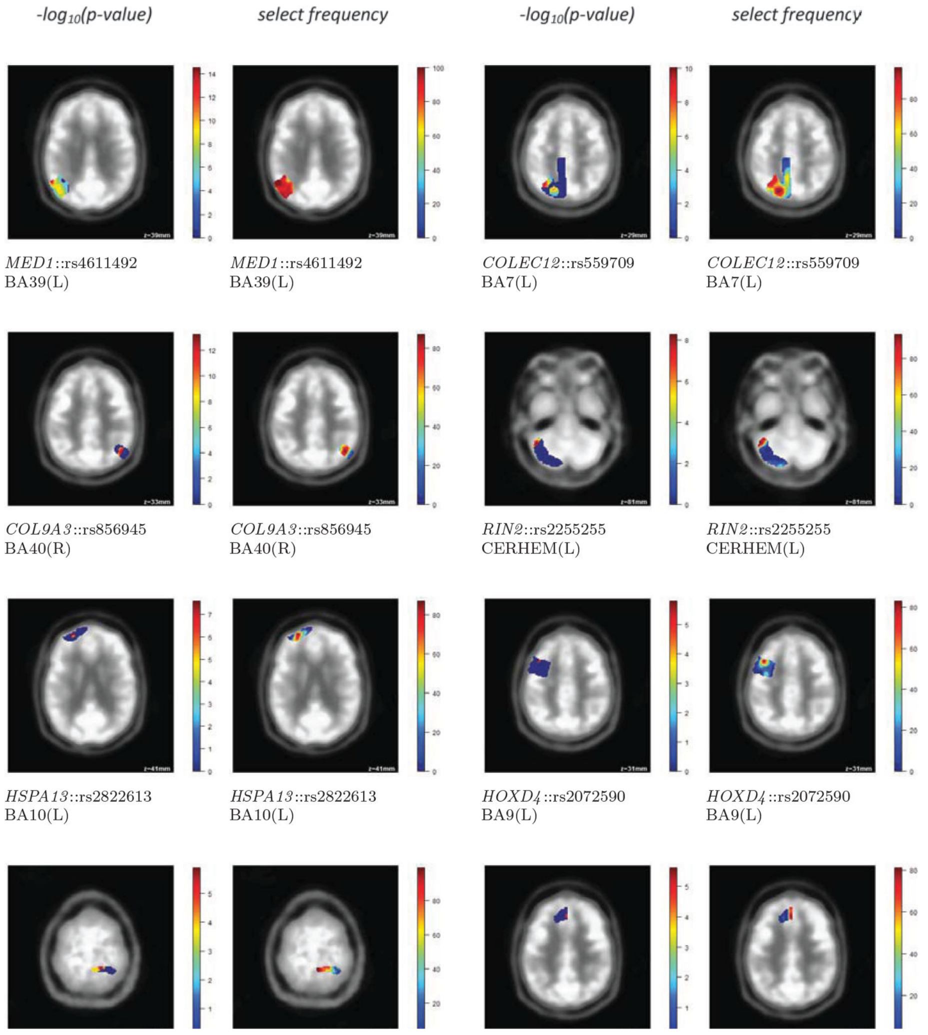
PACS2::rs1008628 BA40(L)

PRIMA1::rs4900194 BA38(R)

PRIMA1::rs4900194 BA38(R)

SNX29::rs10500383 BA21(R)

SNX29::rs10500383 BA21(R)



-log$_{10}$(p-value) select frequency -log$_{10}$(p-value) select frequency

MED1::rs4611492 BA39(L)

MED1::rs4611492 BA39(L)

COLEC12::rs559709 BA7(L)

COLEC12::rs559709 BA7(L)

COL9A3::rs856945 BA40(R)

COL9A3::rs856945 BA40(R)

RIN2::rs2255255 CERHEM(L)

RIN2::rs2255255 CERHEM(L)

HSPA13::rs2822613 BA10(L)

HSPA13::rs2822613 BA10(L)

HOXD4::rs2072590 BA9(L)

HOXD4::rs2072590 BA9(L)

Figure 2:

The most significant SNPs' effects, their $-\log_{10}$(p–values) on voxels across the associated region, and their selective frequency pattern on the region.

Table 1:

Selection results

Correlation structure & Setting		Proposed								Pairwise	
		First stage				Second stage					
		Direct Selection		Stability Selection		Direct Selection		Stability Selection			
	n	**SE(1)**	**SP(1)**	**SE(1)**	**SP(1)**	**SE(2)**	**SP(2)**	**SE(2)**	**SP(2)**	**SE**	**SP**
I	200	0.98 (2e-3)	0.98 (2e-3)	1 (0)	0.98 (1e-3)	0.75 (2e-3)	0.77 (8e-4)	0.82 (3e-3)	0.84 (2e-3)	7e-4 (1e-6)	0.999 (1e-4)
	500	1(0)	1(0)	1(0)	1(0)	0.95 (1e-3)	0.87 (3e-4)	0.98 (3e-4)	0.97 (5e-4)	0.024 (4e-4)	0.999(1e-5)
	1000	1 (0)	1 (0)	1 (0)	1 (0)	0.98 (1e-3)	0.93 (2e-4)	1.00 (7e-5)	0.99 (4e-4)	0.14 (1e-3)	0.999 (3e-5)
II	200	0.98 (2e-3)	0.97 (2e-3)	1 (0)	0.98 (1e-3)	0.74 (2e-3)	0.77 (8e-4)	0.81 (3e-3)	0.83 (2e-3)	7e-4 (1e-4)	0.999 (1e-6)
	500	1 (0)	1 (0)	1 (0)	1 (0)	0.95 (2e-3)	0.86 (4e-4)	0.99 (5e-4)	0.97 (6e-4)	0.024 (4e-4)	0.999 (1e-5)
	1000	1 (0)	1 (0)	1 (0)	1 (0)	0.98 (1e-3)	0.93 (1e-4)	0.99 (1e-4)	0.99 (3e-4)	0.14 (1e-3)	0.999 (2e-5)
III	200	0.98 (2e-3)	0.96 (2e-3)	1 (0)	0.98 (1e-3)	0.74 (2e-3)	0.77 (8e-4)	0.81 (3e-3)	0.83 (2e-3)	7e-4 (1e-4)	0.999 (1e-6)
	500	1 (0)	1 (0)	1 (0)	1 (0)	0.94 (2e-3)	0.87 (3e-4)	0.99 (4e-4)	0.97 (5e-4)	0.024 (4e-4)	0.999 (1e-5)
	1000	1 (0)	1 (0)	1 (0)	1 (0)	0.98 (1e-3)	0.93 (1e-4)	0.99 (1e-4)	0.99 (4e-4)	0.14 (4e-4)	0.999 (3e-5)

• Numbers in parenthesis are empirical standard deviations.

Table 2:

Top selected genes, their associated regions and within them the top selected SNPs

		gene information		top selective SNP in gene		associated	effect type	reference
name	chr	num. SNP in gene	% variance by 20 PCs	SNP name	most sig. p-value	ROI		
IL8RB	2	11	100%	rs6436025	4.8e-12	Superior parietal cortex_BA40(L)	G×AD	Liu et al. [31], Vallées et al. [52], Horuk et al. [23]
					3.8e-11	Superior parietal cortex_BA39(L)	G×AD	Desbaillets et al. [16], Tsai et al. [51], Xia et al. [57]
					2.9e-10	Superior parietal cortex_BA7(L)	G×AD	Wang et al. [54]
					7.0e-09	Superior parietal cortex_BA39(R)	G×AD	
					1.9e-08	Posterior cingulated_BA31(L)	G×AD	
					7.8e-08	Inferior parietal cortex_BA37(L)	G×AD	
				rs4674246	1.2e-07	Medial frontal cortex_BA9(R)	G×AD	
					1.1e-07	Temporal cortex_BA20(R)	G×AD	
CDC42EP3	2	36	98.6%	rs11692666	1.2e-08	Occipital cortex_BA18(R)	G	–
					2.0e-7	Occipital cortex_BA19(R)	G	
AK096399	8	40	98.6%	rs1425765	3.3e-13	Occipital cortex_BA18(R)	G×AD	Cannon et al. [11]
					9.3e-13	Superior parietal cortex_BA39(R)	G×AD	
					3.0e-10	Superior parietal cortex_BA7(R)	G×AD	
				rs269197	5.8e-07	Superior parietal cortex_BA7(L)	G×AD	
SAKCA	10	109	81%	rs16934331	1.6e-11	Superior parietal cortex_BA39(R)	G×AD	Burns et al. [10], Ertekin-Taner [17]
				rs1871066	1.0e-07	Posterior cingulated_BA31(R)	G	
				rs3781141	6.9e-07	Superior parietal cortex_BA39(R)	G×AD	
				rs2247557	8.3e-07	Primary somatosensory cortex_BA2(R)	G×AD	
GALNT4	12	9	100%	rs12426089	3.5e-12	Occipital cortex_BA19(L)	G×AD	–
BC007399	12	36	97%	rs734148	5.1e-14	Superior parietal cortex_BA39(R)	G×AD	–
					1.4e-8	Occipital cortex_BA19(R)	G	
				rs12423428	7.1e-09	Superior parietal cortex_BA39(R)	G	
GJB2	13	25	98.1%	rs4769088	2.6e-10	Superior parietal cortex_BA40(R)	G×AD	Lingala et al. [32], Yaffe et al. [59], Heikaus et al. [21]

		gene information		top selective SNP in gene		associated	effect type	reference
name	chr	num. SNP in gene	% variance by 20 PCs	SNP name	most sig. p-value	ROI		
				rs10870680	1.3e-08	Superior parietal cortex_BA40(R)	G×AD	
				rs945373	2.7e-08	Superior parietal cortex_BA40(R)	G	
PACS2	14	8	100%	rs1008628	1.2e-08	Superior parietal cortex_BA40(L)	G	Aslan et al. [2], Nitsch et al. [39]
PRIMA1	14	48	90.3%	rs4900194	4.9e-08	BA38(R)	G×AD	Xie et al. [58]
				rs12895346	2.9e-07	BA38(R)	G×AD	
				rs2064930	3.4e-07	BA38(R)	G×AD	
SNX29	16	249	76.2%	rs10500383	3.6e-11	Temporal cortex_BA21(R)	G×AD	Teasdale and Collins [50]
				rs10500383	3.0e-09	Temporal cortex_BA22(R)	G×AD	
MED1	17	4	100%	rs4611492	3.1e-15	Superior parietal cortex_BA39(L)	G×AD	Giordano and Macaluso [20] Wong et al. [56]
					8.2e-14	Superior parietal cortex_BA39(R)	G×AD	
					5.6e-13	Temporal cortex_BA22(L)	G×AD	
					5.9e-11	Occipital cortex_BA19(R)	G×AD	
					1.9e-10	Temporal cortex_BA21(L)	G×AD	
COLEC12	18	127	75%	rs559709	9.4e-11	Superior parietal cortex_BA7(L)	G×AD	Nakamura et al. [38]
				rs12960602	2.5e-09	Superior parietal cortex_BA39(L)	G×AD	
COL9A3	20	29	96.6%	rs856945	4.6e-14	Superior parietal cortex_BA40(R)	G×AD	Solovieva et al. [45], Asamura et al. [1]
					6.1e-14	Superior parietal cortex_BA39(R)	G×AD	
					1.2e-10	Superior parietal cortex_BA7(R)	G×AD	
RIN2	20	85	80.4%	rs2255255	5.3e-9	CERHEM(L)	G	–
HSPA13	21	35	99.2%	rs2822613	2.6e-08	Medial frontal cortex_BA10(L)	G×AD	Broer et al. [9]