The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

A structured brain-wide and genome-wide association study using ADNI PET images

Yanming Li[1]*, Bin Nan[2], Ji Zhu[3] and for the Alzheimer's Disease Neuroimaging Initiative

[1]*Department of Biotatistics & Data Science, University of Kansas Medical Center*
[2]*Department of Statistics, University of California at Irvine*
[3]*Department of Statistics, University of Michigan*

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

MSC 2010: Primary 62H20; secondary 62J07

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 methods, 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 resorting to large number of multiple comparisons while effectively controlling the false discoveries. Validity of the proposed approach is demonstrated by both theoretical investigation and numerical simulations. We apply our proposed method to a brain-GWAS using ADNI PET imaging and genomic data. We confirm previously reported association signals and also uncover several novel SNPs and genes that either are associated with brain glucose metabolism or have their association significantly modified by Alzheimer's disease status. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

1. INTRODUCTION

Human brain structures are highly heritable (Braber et al., 2013; Peper et al., 2007). The association patterns between the brain and the genome furnish important information about development and progression mechanisms of chronic cognitive diseases such as Alzheimer's disease (AD) (McKhann et al., 2011). Modern technologies such as the 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 ($\sim$ millions of voxels) and genotype data ($\sim$ millions of SNPs), it is extremely challenging to detect the true association signals immersed in the ultrahigh-dimensional noise. Many current brain-GWAS studies look at a single-voxel-to-single-SNP pair at a time (Stein et al., 2010a). 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 (Stein et al., 2010a; Ge et al., 2012).

Marginal pairwise approaches treat different voxel-to-SNP pairs as independent. A joint model with all voxels and all SNPs considered simultaneously is often of more scientific interest.

* *Author to whom correspondence may be addressed.*
E-mail: yli8@kumc.edu

CJS ???

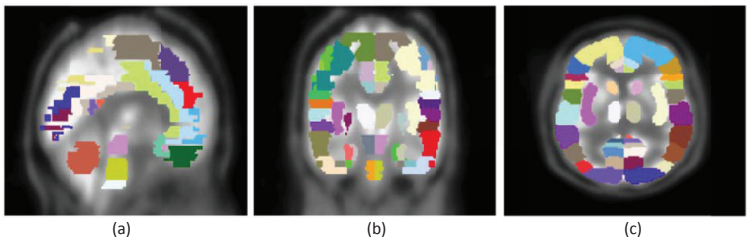


FIGURE 1: : Illustration of mapping Brodmann atlas of ROIs onto segmented PET images. ROIs are highlighted with colors. (a) Sagittal slice at midline. (b) Coronal slice at midline. (c) Axial slice at midline.

Compared to marginal pairwise approaches, joint modelling has enormous potential to improve the power of detecting association signals. Multivariate linear regression is a common technique for jointly modelling 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, Nan, & Zhu (2015) 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 the responses and the predictors. They show that the power to detect the true association signals can be significantly increased by incorporating the grouping structures. However, it is computationally infeasible to fit the MSGLasso directly 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 (Li, Hong, & Li, 2019).

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 (Meinshausen & Bühlmann, 2010) is used in both stages to enhance the stability of the selection and control false positives.

The contribution of our proposed method to variable selection is two-fold. First, it is a joint modelling approach that involves both ultrahigh-dimensional responses and ultrahigh-dimensional predictors. It avoids resorting to a huge number of downstream hypothesis tests and multiple comparisons. Second, it is a structured approach that takes into consideration the grouping structures of both the responses and the predictors. These unique characteristics enable our proposed method to significantly increase the power to identify true signals and, at the same time, to 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 in the brain metabolic pattern as diagnostics of AD progression (Mosconi, 2005). We emphasize that the proposed method is applicable to a wide range of brain-GWAS studies with different imaging modalities or molecular data types, such as functional magnetic resonance imaging (fMRI), methylation, copy number variation and mitochondrial 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 report to conduct a structured brain-GWAS at voxel and SNP levels using a joint model. Compared to the pairwise approaches (Stein et al., 2010a; Ge et al., 2012) and other marginal approaches such as gene-based analysis (Hibar et al., 2011) that regress 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, our proposed method is in general more efficient compared to the pairwise approaches (Stein et al., 2010a). 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 background prior to 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 linking ultrahigh-dimensional neuroimaging responses and genetic DNA predictors.

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

$$\begin{aligned}\boldsymbol{Y} = \mathbf{I}\boldsymbol{\beta}_0^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{B}_X + \mathbf{I}_{ad}\boldsymbol{\beta}_{ad}^{\mathrm{T}} + \mathbf{I}_{mci}\boldsymbol{\beta}_{mci}^{\mathrm{T}} + (\boldsymbol{X} \times \mathbf{I}_{ad})\boldsymbol{B}_{Xad} + (\boldsymbol{X} \times \mathbf{I}_{mci})\boldsymbol{B}_{Xmci} + \\ \mathbf{Age}\boldsymbol{\beta}_{age}^{\mathrm{T}} + \mathbf{Sex}\boldsymbol{\beta}_{sex}^{\mathrm{T}} + \mathbf{E}, \qquad (1)\end{aligned}$$

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, and **Age** and **Sex** are length-n covariate vectors encoding age and sex, respectively. Here $\boldsymbol{\beta}_0$ is a length-Q grand intercept vector; $\boldsymbol{\beta}_{ad} = \mathbf{I}_Q\beta_{ad}$, $\boldsymbol{\beta}_{mci} = \mathbf{I}_Q\beta_{mci}$, $\boldsymbol{\beta}_{age} = \mathbf{I}_Q\beta_{age}$, and $\boldsymbol{\beta}_{sex} = \mathbf{I}_Q\beta_{sex}$ are coefficient vectors for AD indicator, MCI indicator, age and sex, respectively, where $\mathbf{I}_Q$ is a length-Q vector with entries 1; $\boldsymbol{B}$, $\boldsymbol{B}_{Xad}$, and $\boldsymbol{B}_{Xmci}$ are $P \times Q$ regression coefficient matrices for genetic, genetic-AD interaction and genetic-MCI interaction effects, respectively. The symbol $\mathbf{E}$ represents 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 the variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered, $\boldsymbol{\beta}_0$ is zero and the model specified in Equation (1) reduces to

$$\boldsymbol{Y} = \mathbb{X}\mathbb{B} + \mathbf{E} \qquad (2)$$

with $\mathbb{X} = (\boldsymbol{X}, \mathbf{I}_{ad}, \mathbf{I}_{mci}, \boldsymbol{X} \times \mathbf{I}_{ad}, \boldsymbol{X} \times \mathbf{I}_{mci}, \mathbf{Age}, \mathbf{Sex})$ being the grand predictor matrix and $\mathbb{B} = (\boldsymbol{B}_X^{\mathrm{T}}, \boldsymbol{\beta}_{ad}, \boldsymbol{\beta}_{mci}, \boldsymbol{B}_{Xad}^{\mathrm{T}}, \boldsymbol{B}_{Xmci}^{\mathrm{T}}, \boldsymbol{\beta}_{age}, \boldsymbol{\beta}_{sex})^{\mathrm{T}}$ 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 corresponding 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 induce a block grouping structure on $\boldsymbol{B}_X$, with row blocks corresponding to gene groups and column blocks corresponding to ROI groups. These same groups 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 the 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 small number of SNPs (SNP-disease interactions) compared to the sample size. In the following analyses, we assume that the variables $\mathbf{I}_{ad}$, $\mathbf{I}_{mci}$, **Age** and **Sex** belonging to the model specified in Equation (2) each form a group in their own right.

2.1. First stage: Selecting important ROI-to-gene blocks

In the first stage, we use the multivariate group lasso (Li, Nan, & Zhu, 2015; Yuan & Lin, 2006) to select the important ROI-to-gene pairs. This stage serves as a screening step, ruling out the unimportant ROI-to-gene pairs by shrinking the corresponding association blocks to zero. To reduce the dimensionality of the 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 evidence of the associations between their representative ROIs and genes. The advantage of using the PCs is two-fold. First, it helps to reduce the input dimensionality while keep the grouping structure and essential information within each group, and therefore improves the efficiency of group-level selection. Second, since PCs are orthogonal (independent) to each other, using them avoid the complications arising from collinearity between predictors or from overlapping grouping structures, since genes are often overlapping with each other.

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

$$\arg\min_{\mathbf{\Gamma}} \frac{1}{2n} \|\mathbb{P}_Y - \mathbb{P}_X \mathbf{\Gamma}\|_2^2 + \lambda_1 \sum_{rg \in \mathcal{R} \otimes \mathcal{G}} \omega_{rg}^{1/2} \|\mathbf{\Gamma}_{rg}\|_2, \tag{3}$$

where $\|\cdot\|_2$ denotes the l_2 norm. Here $\mathbf{\Gamma}$ is the regression coefficient matrix between the PC matrices and $\mathbf{\Gamma}_{rg}$ is a submatrix block between rth ROI and gth gene group. The group lasso penalty $\sum_{rg \in \mathcal{R} \otimes \mathcal{G}} \omega_{rg}^{1/2} \|\mathbf{\Gamma}_{rg}\|_2$ aims to shrink the unimportant $\mathbf{\Gamma}_{rg}$ blocks to zero, and ω_{rg} is a non-negative weight assigned to $\mathbf{\Gamma}_{rg}$, $r = 1, \ldots, R$, $g = 1, \ldots, G$. In our brain-wide GWAS, we use $\omega_{rg} = \sqrt{v \times s}$ (Yuan & Lin, 2006; Silver, Montana, & ADNI, 2012), where v is the number of PCs in the rth ROI group and s is the number of PCs in the gth gene group. We set $\omega_{rg} = 0$ if we do not want to penalize the group with label rg. 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 $\mathbf{\Gamma}_{rg}$ selected at the first stage, the corresponding ROI-to-gene pairs are passed to the second stage. In the second stage, we narrow our focus to the associations for those same pairs at voxle-to-SNP levels. For each selected ROI-to-gene pair, we solve the following multivariate lasso problem (Li, Nan, & Zhu, 2015; Kohannim et al., 2012; Friedman, Hastie, & Tibshirani, 2010),

$$\arg\min_{\boldsymbol{B}_{rg}} \frac{1}{2n}\|\boldsymbol{Y}_r - \boldsymbol{X}_g\boldsymbol{B}_{rg}\|_2^2 + \lambda_2 \sum_{\beta_{jk}\in\boldsymbol{B}_{rg}} \omega_{jk}|\beta_{jk}|, \tag{4}$$

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) belonging to 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 nonnegative weight corresponding to β_{jk}. If $\omega_{jk} = 0$, then β_{jk} will not be penalized. In our ANDI data analysis, we set $\omega_{jk} = 1$ for all β_{jk}s that corresponds to either a SNP main effect or a SNP-to-disease interaction effect.

2.3. Stability selection and control of false discoveries

Stability selection (Meinshausen & Bühlmann, 2010) is employed in both stages. We fit the models identified in Equations (3) and (4) multiple times, say K, 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 frequency of being selected among the total K times of selections exceeds a certain specified threshold.

The advantages of stability selection are three-fold. First, it can reduce the random variation in the data that arises from sampling or measurement error. Second, it saves the computing cost associated with choosing the tuning parameters λ_1 and λ_2. Instead of using cross-validation to select optimal tuning parameters, stability selection prescribes using a fixed set of tuning parameter values on re-randomized datasets. As long as the proposed fixed tuning parameter values belong to a reasonable range, i.e., they are neither too large so that they shrink almost everything to zeros nor too small so that they barely shrink anything, the corresponding variable selection results are quite stable. Figure S.3 in the online Supplementary Materials illustrates that the top signals identified in the analysis of the ADNI PET imaging and genetic data are robustly selected when using bootstrapped samples and different values of the tuning parameters. Stability selection can be easily implemented and run on multi-core computing clusters and therefore is much more efficient computationally. Third, stability selection provides a quantitative way to govern the number of false discoveries, an issue that 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 (Bickel, Ritov, & Tsybakov, 2009).

First, we introduce some notation. Let $\mathcal{J}_1(\mathbb{B}) = \{jk : |\beta_{jk}| \neq 0\}$ be the index set of nonzero elements in $\mathbb{B}$, and let $\mathcal{J}_2(\mathbb{B}) = \{rg \in \mathcal{R}\otimes\mathcal{G}, \|\boldsymbol{B}_{rg}\|_2 \neq 0\}$ be the index set of nonzero groups. Define $M_1(\mathbb{B}) = \sum_{jk} I(\beta_{jk} \neq 0) = |\mathcal{J}_1(\mathbb{B})|$ and $M_2(\mathbb{B}) = \sum_{rg\in\mathcal{R}\otimes\mathcal{G}} I(\|\boldsymbol{B}_{rg}\|_2 \neq 0) = |\mathcal{J}_2(\mathbb{B})|$. 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 summarized in Theorem 1.

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

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

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

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

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

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

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

$$r_1 = \left(\frac{nc^*\sqrt{d^*}\eta_1}{\lambda_1 d_* M_2}\right)^{1/2}, r_2 = \left(\frac{nc^*\eta_2^2}{\lambda_1^2 d_* M_2}\right)^{1/2}, \bar{c} = c^*/c_* \text{ and}$$

$$C_2 = 2 + 4r_1^2 + 4\sqrt{d\bar{c}}r_2 + 4d\bar{c}.$$

Let $\sigma^* = \sigma\sqrt{\max_{g\in\mathcal{G}} p_g}$. Assume that the tunning parameter, λ_1, in the model specified in Equation (3) satisfies

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

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 $N_d = \sum_{rg\in\mathcal{R}\otimes\mathcal{G}} \omega_{rg}$, $c_0 \geq 0$ 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 $\inf\emptyset = \infty$. Here $a \vee b = \max\{a, b\}$.

Let $\bar{q} = \max\{q_1,\ldots,q_R\}$ and $\bar{p} = \max\{p_1,\ldots,p_G\}$. Assume that the tuning parameter, λ_2, in the model specified in Equation (4) satisfies

$$\lambda_2 = 2\sigma A\{\log(\bar{q}\bar{p})/n\}^{1/2}$$

for some constant $A > \sqrt{2}$.

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 $\mathbb{X}^{\mathrm{T}}\mathbb{X}/n$ and $M_1^*(\mathbb{B}^*) = \max_{rg\in\mathcal{R}\otimes\mathcal{G}} M_1(\boldsymbol{B}_{rg}^*)$. Assume gGSC, SRC, iREC and the conditions specified in (iv) hold. Then with probability converging to 1 as $n \to \infty$, we have the following oracle selection bounds for group- and individual-level signals:*

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

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

When gGSC, SRC and the conditions specified in (iv) hold, Wei & Huang (2010) showed that the group-level selection bound holds for the univariate-response group lasso. The proof of the inequality specified in Equation (6) follows Theorem 2.1 in (Wei & Huang, 2010), 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 $\mathbb{P}_{X,\mathcal{A}_1} = \mathbb{X}_{\mathcal{A}_1}\mathbf{W}$, 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 $\|\mathbb{P}_{X,\mathcal{A}_1}\boldsymbol{\nu}\|_2^2/\{n\|\boldsymbol{\nu}\|_2^2\} = \|\mathbb{X}_{\mathcal{A}_1}\mathbf{W}\boldsymbol{\nu}\|_2^2/\{n\|\boldsymbol{\nu}\|_2^2\} = \|\mathbb{X}_{\mathcal{A}_1}\mathbf{W}\boldsymbol{\nu}\|_2^2/\{n\boldsymbol{\nu}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\boldsymbol{\nu}\} = \|\mathbb{X}_{\mathcal{A}_1}\boldsymbol{\nu}'\|_2^2/\{n\|\boldsymbol{\nu}'\|_2^2\}$, 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 identified in Equation (7) follows directly from the bound indicated in Equation (6) and the multivariate lasso oracle selection bound introduced in Theorem 2 in Li, Nan, & Zhu (2015).

3. A SIMULATION STUDY

We investigated the empirical selection performance for our proposed two-stage method via 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 were set to be zeros. Half of the coefficients within diagonal blocks were randomly generated from Unif$([-5,-3]\cup[3,5])$ and the other half were set to equal zero (therefore, the sparsity within important coefficient blocks was 0.5). Once $\mathbb{B}$ was generated, it remained fixed in all the experiments.

We assumed the $\mathbb{X}$ groups were uncorrelated. Within-group $\mathbb{X}$ variables were generated from a multivariate 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 generated the noise variables $\mathbf{E}$ from a multivariate normal distribution with one of the following three correlation structures and unit marginal variances:

I. Independent $\boldsymbol{Y}$ groups: The variables within each $\boldsymbol{Y}$ group followed an AR1(0.5) correlation structure.
II. Weakly correlated $\boldsymbol{Y}$ groups: The variables within each $\boldsymbol{Y}$ group followed an AR1(0.5) correlation structure. The variables from different $\boldsymbol{Y}$ groups were 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 was CS$(0.1)\otimes$AR1(0.5), where $\otimes$ is the Kronecker product.
III. Moderately correlated $\boldsymbol{Y}$ groups: The variables within each $\boldsymbol{Y}$ group followed an AR1(0.5) correlation structure. The variables from different $\boldsymbol{Y}$ groups followed a CS(0.5) correlation structure. The overall $\boldsymbol{Y}$ correlation structure was CS$(0.5)\otimes$AR1(0.5).

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

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

correspond to

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

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

where the superscript $*$ indicates the true values. The second stage individual-level SE and SP equal

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

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

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

The simulation results demonstrate 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 was far more powerful than the pairwise approach. Especially for the first-stage group-level selection, our approach provided almost perfect selection performance even when the sample size is very small. For the second-stage individual-level selection, the selection performance improved significantly as the sample size increased. The selection performance was similar across all three different correlation structures for the simulated responses.

4. ANALYSIS OF THE ADNI FDG-PET AND SNP DATA

The ADNI data used in our structured brain-GWAS analysis consists of 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 involve 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 the model specified in Equation (1) to the ADNI PET imaging and genomic data using our proposed method.

4.1. PET images and ROI's

The 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 were 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 were included in our study. Each image contains 349,182 voxels embedded in a $160\times160\times96$ 3D array. All these images were preprocessed to produce a uniform isotropic resolution.

To incorporate the brain anatomic structures, the PET images were segmented according to the Brodmann atlas (Brodmann, 2010). As a result, the voxels in each image were grouped

TABLE 1: : Selection results for the simulation study; the numbers in parenthesis are empirical standard deviations.

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	0.98	1	0.98	0.75	0.77	0.82	0.84	7e-4	0.999
		(2e-3)	(2e-3)	(0)	(1e-3)	(2e-3)	(8e-4)	(3e-3)	(2e-3)	(1e-4)	(1e-6)
	500	1	1	1	1	0.95	0.87	0.98	0.97	0.024	0.999
		(0)	(0)	(0)	(0)	(1e-3)	(3e-4)	(3e-4)	(5e-4)	(4e-4)	(1e-5)
	1000	1	1	1	1	0.98	0.93	1.00	0.99	0.14	0.999
		(0)	(0)	(0)	(0)	(1e-3)	(2e-4)	(7e-5)	(4e-4)	(1e-3)	(3e-5)
II	200	0.98	0.97	1	0.98	0.74	0.77	0.81	0.83	7e-4	0.999
		(2e-3)	(2e-3)	(0)	(1e-3)	(2e-3)	(8e-4)	(3e-3)	(2e-3)	(1e-4)	(1e-6)
	500	1	1	1	1	0.95	0.86	0.99	0.97	0.024	0.999
		(0)	(0)	(0)	(0)	(2e-3)	(4e-4)	(5e-4)	(6e-4)	(4e-4)	(1e-5)
	1000	1	1	1	1	0.98	0.93	0.99	0.99	0.14	0.999
		(0)	(0)	(0)	(0)	(1e-3)	(1e-4)	(1e-4)	(3e-4)	(1e-3)	(2e-5)
III	200	0.98	0.96	1	0.98	0.74	0.77	0.81	0.83	7e-4	0.999
		(2e-3)	(2e-3)	(0)	(1e-3)	(2e-3)	(8e-4)	(3e-3)	(2e-3)	(1e-4)	(1e-6)
	500	1	1	1	1	0.94	0.87	0.99	0.97	0.024	0.999
		(0)	(0)	(0)	(0)	(2e-3)	(3e-4)	(4e-4)	(5e-4)	(4e-4)	(1e-5)
	1000	1	1	1	1	0.98	0.93	0.99	0.99	0.14	0.999
		(0)	(0)	(0)	(0)	(1e-3)	(1e-4)	(1e-4)	(4e-4)	(4e-4)	(3e-5)

into 106 Brodmann ROIs. Voxels not indexed by the Brodmann atlas were not considered in the analysis. The regions on the left hemisphere are a symmetric mirror reflection of the ones located 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 found on the right hemisphere.

4.2. Genotypes

The ADNI SNP data were genotyped using an Illumina 610 Quad array with more than 620,000 tag SNPs. Genotyping was performed by Polymorphic DNA Technologies. We grouped the 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 both upstream and downstream. 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 a 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 used the first five PCs in each brain ROI and the first twenty PCs or the first several PCs that explained at least 80% of the variation, whichever was 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 their variations explained by no more than 20 PCs. For example, only seven out of 800 genes on chromosome 20 have less than 60% of their variations explained by their first 20 PCs. Figure S.1 in the online Supplementary Materials 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 *APOE* gene (*APOE-ϵ4*) is the most common genetic risk factor for AD (Corder et al., 2004; Strittmatter et al., 1993). However, the ADNI genetic dataset does not contain the genotypes for the SNPs in the *APOE* gene. We extracted the *APOE-ϵ4* allele information score from the ADNI clinical data and combined it with the first 20 PCs on chromosome 19.

We used the R package MSGLasso (Li, Nan, & Zhu, 2016) to run the multivariate group lasso on the PC matrices. Stability selection (Meinshausen & Bühlmann, 2010) was then carried out on 100 bootstrapped datasets. ROI-to-gene pairs with a stability selection frequency of at least 75% were selected as important ROIs and genes in the first selection stage. For the *APOE* gene, we used *APOE-ϵ4* allele score to fit the model specified in Equation (4) wherever *APOE* was selected.

Meinshausen & Bühlmann (2010) 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} denotes the thresholding frequency used for the selection, which in our case was 75% for the first stage and 80% for the second stage, and q represents the average number of selected variables. In our study, the typical numbers of selected variables ranged from tens to hundreds out of tens of thousands of variables in total, which yielded $q^2/P \ll 1$. 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 used the MSGLasso (Li, Nan, & Zhu, 2016) to fit a multivariate lasso regression on each of the selected ROI-to-gene pairs. Stability selection was then carried out on 100 bootstrapped datasets for each ROI-to-gene pair. Voxel-to-SNP pairs with a selection frequency greater than 80% were selected to be the important individual-level signals. Then we applied 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 ranged from a few to several dozens, which is much smaller than the sample size.

In both stages, we did 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 having a p-value less than 10^{-6} and a selection frequency exceeding 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 that satisfies both criteria, we provide a list of top MCI interactions in Table S.1 in the online Supplementary Materials. Table S.2 lists the top selected ROIs and voxels therein for the *APOE-ϵ4* effects.

Some brain regions are identified as having 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 were found to be 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 pairwise brain-GWAS studies (Stein et al., 2010a). We have confirmed some brain regions associated with genetics that have appeared in the existing literature. For example, Mills et al. (2013) reported associations between lipid metabolism in superior parietal cortices and alternatively spliced isoforms in RNA transcriptome. Other identified regions that were 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 (Braskie, Ringman, & Thompson, 2011) and BA20(R), BA20(L), BA21(R), BA21(L), BA22(R) and BA22(L) in temporal cortices (Stein et al., 2010b; Risacher et al., 2009; Braskie, Ringman, & Thompson, 2011).

Some genetic findings identified in previous studies were confirmed by our brain-GWAS. For example, Wang et al. (2013) found that inhibiting *IL8RB* (*CXCR2*) can turn down amyloid-β production and protect neural cells. Nakamura et al. (2006) found a similar effect for the *COLEC12* (*SRCL*) gene in AD samples. Other direct supports involving AD interactions include Burns et al. (2011) on *SAKCA* (*KCNMA1*), Xie et al. (2010) on *PRIMA*, Nakamura et al. (2006) on *COLEC12* and Broer et al. (2011) on *HSPA13*.

Some gene-to-AD interactions have also been identified in the literature as associated with other cognitive-related diseases such as autism and hearing impairment. Such genes include *AK096399* (Cannon et al., 2010), *GJB2* (Lingala, Sankarathi, & Penagaluru, 2009), *SNX29* (Teasdale & Collins, 2012), *MED1* (Giordano & Macaluso, 2011; Wong et al., 2013) and *COL9A3* (Solovieva et al., 2006; Asamura et al., 2005).

We also confirmed some gene effects on brain metabolizing. For example, *CDC42EP3* encodes a certain family of guanosine triphosphate metabolizing proteins and the gene is weakly expressed in the brain. *PACS2* plays a role in membrane traffic with tumour-necrosis-factor-related apoptosis-inducing-ligand (TRAIL) induced apophasis (Aslan et al., 2009), which in turn can cause human brain cell death (Nitsch et al., 2000).

Our findings also provided evidence about indirect genetic effects on certain chemical compounds or protein translocation, which are reflected in the PET scans and may be associated with AD. For example, Dai et al. (2013) and Sakamoto & Holman (2008) demonstrated that *TBC1D4* plays a role in regulation of *GluT4* traffic, which, on the other hand is associated with AD (Talbot et al., 2012; Yang, Li, & Liu, 2013). Nolte et al. (2006) and Lu, He, & Zhong (2007) have given a chain of relationships of *HOXD4* gene to *PAX6* protein to AD.

There are also several novel signals which have not previously been reported in the 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 that of the pairwise approaches (Stein et al., 2010a; Ge et al., 2012), as our method 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 reduce the computational time, we parallelized the computational jobs on multi-core computing clusters. Also, our approach has more power due to the integration of the brain and genome grouping structures. In Stein et al. (2010a), no significant voxel-to-SNP signals were found due to the huge number of multiple comparisons that were carried out.

We recognize that post-selection inference is biased. Simultaneous selection, estimation and inference have been studied recently (van de Geer et al., 2014; Berk et al., 2013). Kuchibhotla et al. (2020) also provide an upper bound for post-selection inference p-values when taking into account the selection bias. These enhancements of our proposed method will be investigated in future studies.

ACKNOWLEDGEMENTS

The authors would like to thank the two reviewers and Dr. Linglong Kong for helpful suggestions. The authors would also like to thank Dr. David Matthews for proofreading the manuscript. 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.

BIBLIOGRAPHY

Asamura, K., Abe, S., Fukuoka, H., Nakamura, Y., & 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.

Aslan, J., You, H., Williamson, D. M., Endig, J., amd L Thomas, R. Y., Shu, H., Du, Y., Milewski, R., Brush, M., Possemato, A., Sprott, K., Fu, H., Greis, K., Runckel, D., Vogel, A., & 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.

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

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Danzig selector. *Ann. Statist.*, 37, 1705–1732.

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

Braskie, N., Ringman, J., & Thompson, P. (2011). Neuroimaging measures as endophenotypes in Alzheimer's disease. *International Journal of Alzheimer's Disease*. http://dx.doi.org/10.4061/2011/490140.

Brodmann, K. (2010). *Brodmann's Localisation in the Cerebral Cortex*. Springer Science+Business Media, New York City.

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., & van Duijn, C. (2011). Association of *HSP70* and its co-chaperones with Alzheimer's disease. *J Alzheimers Dis.*, 25(1), 93–102.

Burns, L., Minster, R., Demirci, F., Barmada, M., Ganguli, M., Lopez, O. L., DeKosky, S., & 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.

Cannon, D., Miller, J., Robison, R., Villalobos, M., Wahmhoff, N., Allen-Brady, K., McMahon, W., & Coon, H. (2010). Genome-wide linkage analyses of two repetitive behavior phenotypes in utah pedigrees with autism spectrum disorders. *Molecular Autism*, 1(1), 3.

Corder, E. H., Ghebremedhin, E., Taylor, M. G., Thal, D. R., Ohm, T. G., & 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.

Dai, M., Freeman, B., Shikani, H. J., Bruno, F. P., Collado, J. E., Macias, R., Reznik, S. E., Davies, P., Spray, D. C., Tanowitz, H. B., Weiss, L. M., & Desruisseaux, M. S. (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.

Desbaillets, I., Diserens, A., Tribolet, N., Hamou, M., & Meir, E. V. (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.

Ertekin-Taner, N. (2010). Genetics of Alzheimer disease in the pre- and post-GWAS era. *Alzheimer's Research & Therapy*, 2(1), 3.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., & Nichols, T. E. (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.

Giordano, A. & Macaluso, M. (2011). *Cancer Epigenetics: Biomolecular Therapeutics in Human Cancer*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Heikaus, S., Winterhager, E., Traub, O., & 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.

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

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

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

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., & Zhao, L. (2020). Valid Post-selection Inference in Assumption-lean Linear Regression. *Annals of Statistics*, 48 (5), 2953–2981

Li, Y., Hong, H. G., & Li, Y. (2019). Multiclass linear discriminant analysis with ultrahigh-dimensional features. *Biometrics*, 75(4), 1086–1097.

Li, Y., Nan, B., & Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2), 354–363.

Li, Y., Nan, B., & 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.

Liu, Y., Guo, D., Tian, L., Shang, D., Zhao, W., Li, B., Fang, W., Zhu, L., & 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.

Lingala, H. B., Sankarathi, & Penagaluru, P. R. (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.

Lu, Y., He, X., & 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.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, Jr C. R., Kawas, C. H., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Insti-

tute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease *Alzheimer's & Dementia*, 7(3), 263–269.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B.*, 72, 417–473.

Mills, J., Nalpathamkalam, T., Jacobs, H., Merico, C. J. D., Hu, P., & 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.

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.

Nakamura, K., Ohya, W., Funakoshi, H., Sakaguchi, G., Kato, A., Takeda, M., Kudo, T., & 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.

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

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

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

Risacher, S., West, A. S. J., Shen, L., Firpi, H., & McDonald, B. (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res.*, 6(4), 347–361.

Sakamoto, K. & Holman, G. D. (2008). Emerging role for *AS160/TBC1D4* and *TBC1D1* in the regulation of *GLUT4* traffic. *Am J Physiol Endocrinol Metab*, 295, e29–37.

Silver, M., Montana, G., & ADNI (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.

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

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., & ADNI (2010). Voxelwise genome-wide association study (vgwas). *Neuroimage*, 53(3), 1160–1174.

Stein, J., Hua, X., Morra, J., Lee, S., Hibar, D., Ho, A., Leow, A., Toga, A., Sul, J., Kang, H., Eskin, E., AJ, A. S., Shen, L., Foroud, T., Pankratz, N., Huentelman, M., Craig, D., Gerber, J., Allen, A., Corneveaux, J., DA, D. S., Webster, J., DeChairo, B., Potkin, S., Jack, C., Weiner, M., Thompson, P., & 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.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & 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.

Talbot, K., Wang, H., Kazi, H., Han, L., Bakshi, K. P., Stucky, A., Fuino, R. L., Kawaguchi, K. R., Samoyedny, A. J., Wilson, R. S., Arvanitakis, Z., Schneider, J. A., Wolf, B. A., Bennett, D. A., Trojanowski, J. Q., & Arnold1, S. E. (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.

Teasdale, R. D. & Collins, B. M. (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.

Tsai, H., Frost, E., To, V., Ffrench-Constant, S. R. C., Geertman, R., Ransohoff, R., & 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.

Vallès, A., Grijpink-Ongering, L., de Bree, F., Tuinstra, T., & 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.

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

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

Wei, F. R. & Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4), 1369–1384.

Wong, C. C. Y., Meaburn, E. L., Ronald, A., Price, T. S., Jeffries, A. R., Schalkwyk, L. C., Plomin, R., & Mill, J. (2013). Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. *Molecular Psychiatry*.

Xia, M., Qin, S., McNamara, M., Mackay, C., & Hyman, B. T. (1997). Interleukin-8 receptor B immunoreactivity in brain and neuritic plaques of Alzheimer's disease. *Am J Pathol.*, 150(4), 1267–1274.

Xie, H., Liang, D., Leung, K., Chen, V., Zhu, K., Chan, W., Choi, R., Massoulié, J., & 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.

Yaffe, K., Krueger, K., Cummings, S. R., Blackwell, T., Henderson, V. W., Sarkar, S., Ensrud, K., & 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.

Yang, J., Li, S., & 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.

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

Received 9 July 2009
Accepted 8 July 2010

TABLE 2: : Top selected genes, their associated regions and within them the top selected SNPs.

name	gene information chr	num. SNP in gene	% variance by 20 PCs	top selective SNP in gene SNP name	most sig. p-value	associated ROI	effect type	reference
IL8RB	2	11	100%	rs6436025	4.8e-12	Superior parietal cortex_BA40(L)	G×AD	Liu et al. (2010), Vallès et al. (2006), Horuk et al. (1997)
					3.8e-11	Superior parietal cortex_BA39(L)	G×AD	Desbaillets et al. (1997), Tsai et al. (2002), Xia et al. (1997)
					2.9e-10	Superior parietal cortex_BA7(L)	G×AD	Wang et al. (2013)
					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	
					1.9e-07	Temporal cortex_BA20(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. (2010)
					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	

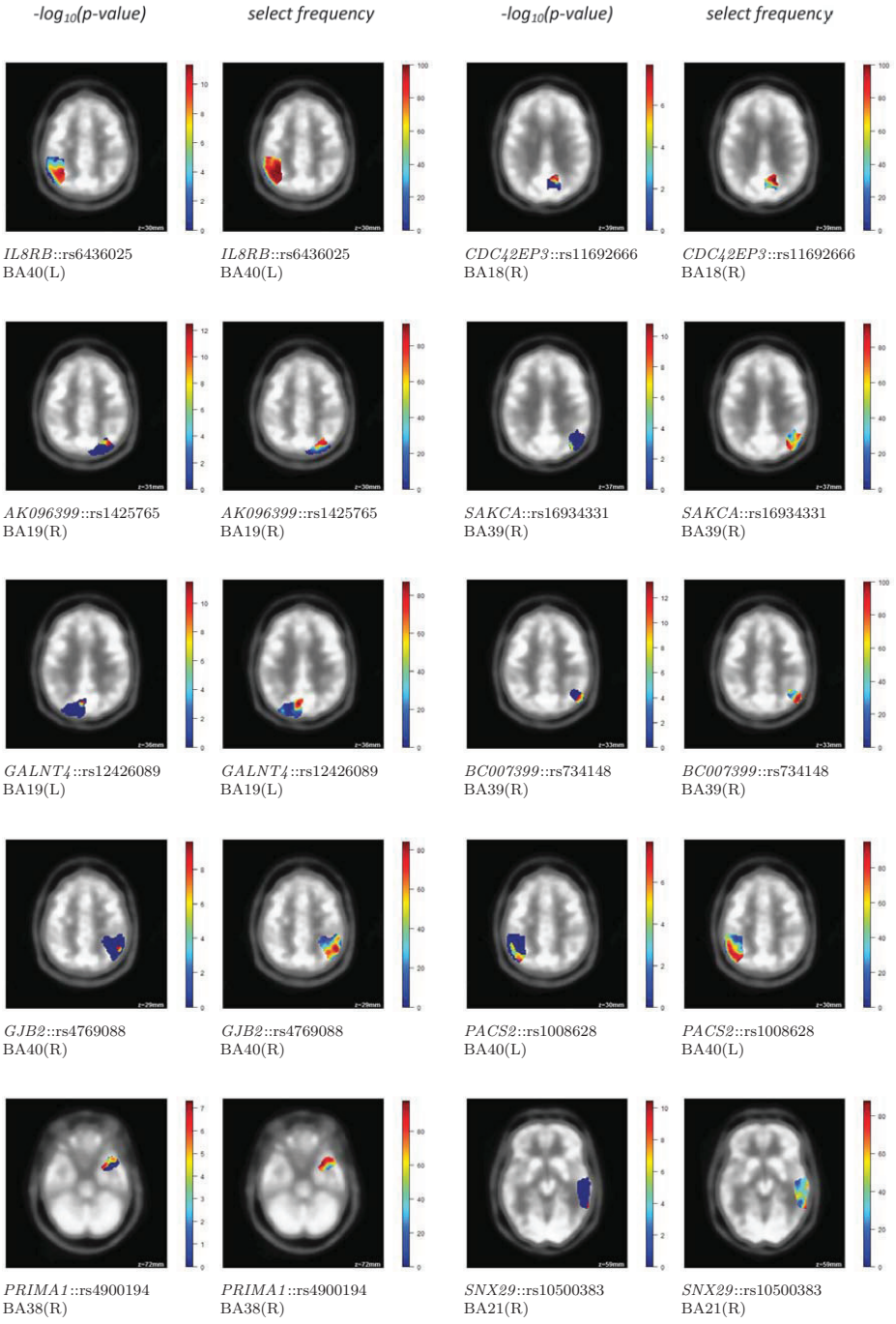
Continued on next page

name	gene information chr	num. SNP in gene	% variance by 20 PCs	top selective SNP in gene SNP name	most sig. p-value	associated ROI	effect type	reference
SAKCA	10	109	81%	rs16934331	1.6e-11	Superior parietal cortex_BA39(R)	G×AD	Burns et al. (2011), Ertekin-Taner (2010)
				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, Sankarathi, & Penagaluru (2009), Yaffe et al. (2005), Heikaus et al. (2002)
				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. (2009), Nitsch et al. (2000)
PRIMA1	14	48	90.3%	rs4900194	4.9e-08	BA38(R)	G×AD	Xie et al. (2010)
				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 & Collins (2012)
				rs10500383	3.0e-09	Temporal cortex_BA22(R)	G×AD	
				rs11859327	8.4e-07	Occipital cortex_BA19(R)	G	

Continued on next page

name	gene information chr	num. SNP in gene	% variance by 20 PCs	top selective SNP in gene SNP name	most sig. p-value	associated ROI	effect type	reference
MED1	17	4	100%	rs4611492	3.1e-15	Superior parietal cortex_BA39(L)	G×AD	Giordano & Macaluso (2011) Wong et al. (2013)
					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. (2006)
				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. (2006), Asamura et al. (2005)
					6.1e-14	Superior parietal cortex_BA39(R)	G×AD	
					1.2e-10	Superior parietal cortex_BA7(R)	G×AD	
					1.7e-9	Occipital cortex_BA19(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. (2011)

DOI:



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)

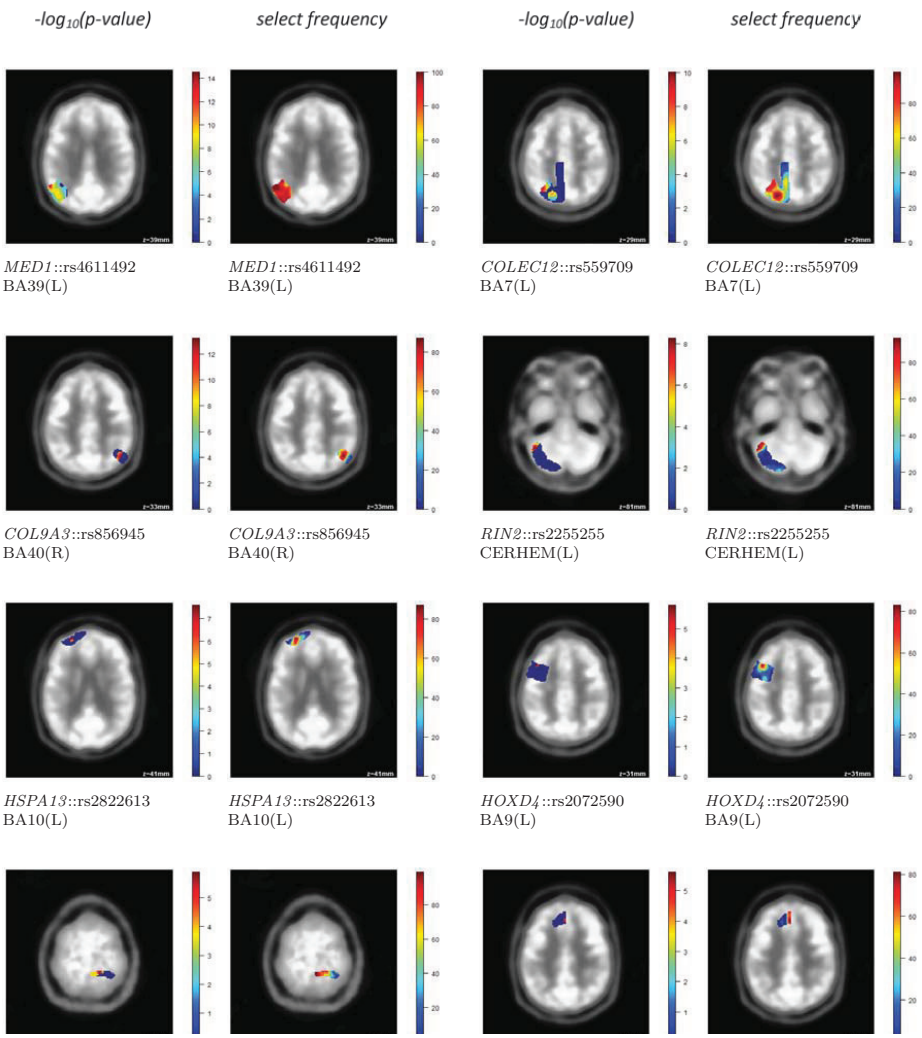
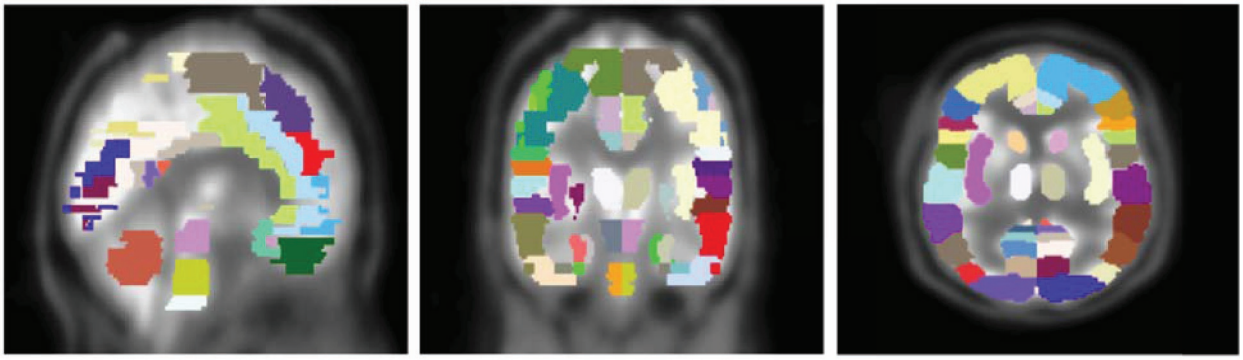
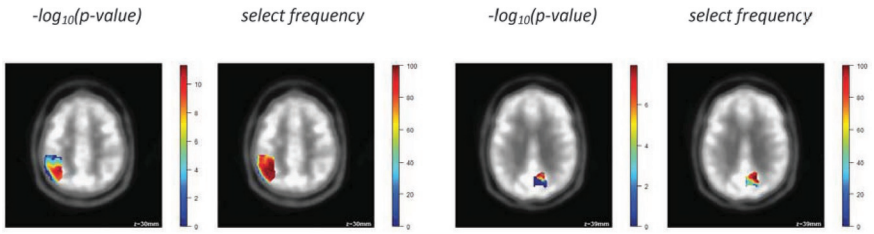


FIGURE 2: : The most significant SNPs' effects, their $-\log_{10}(p-\text{values})$ on voxels across the associated region, and their selective frequency pattern on the region.



(a) (b) (c)

cjs_11605_figure1.eps



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)

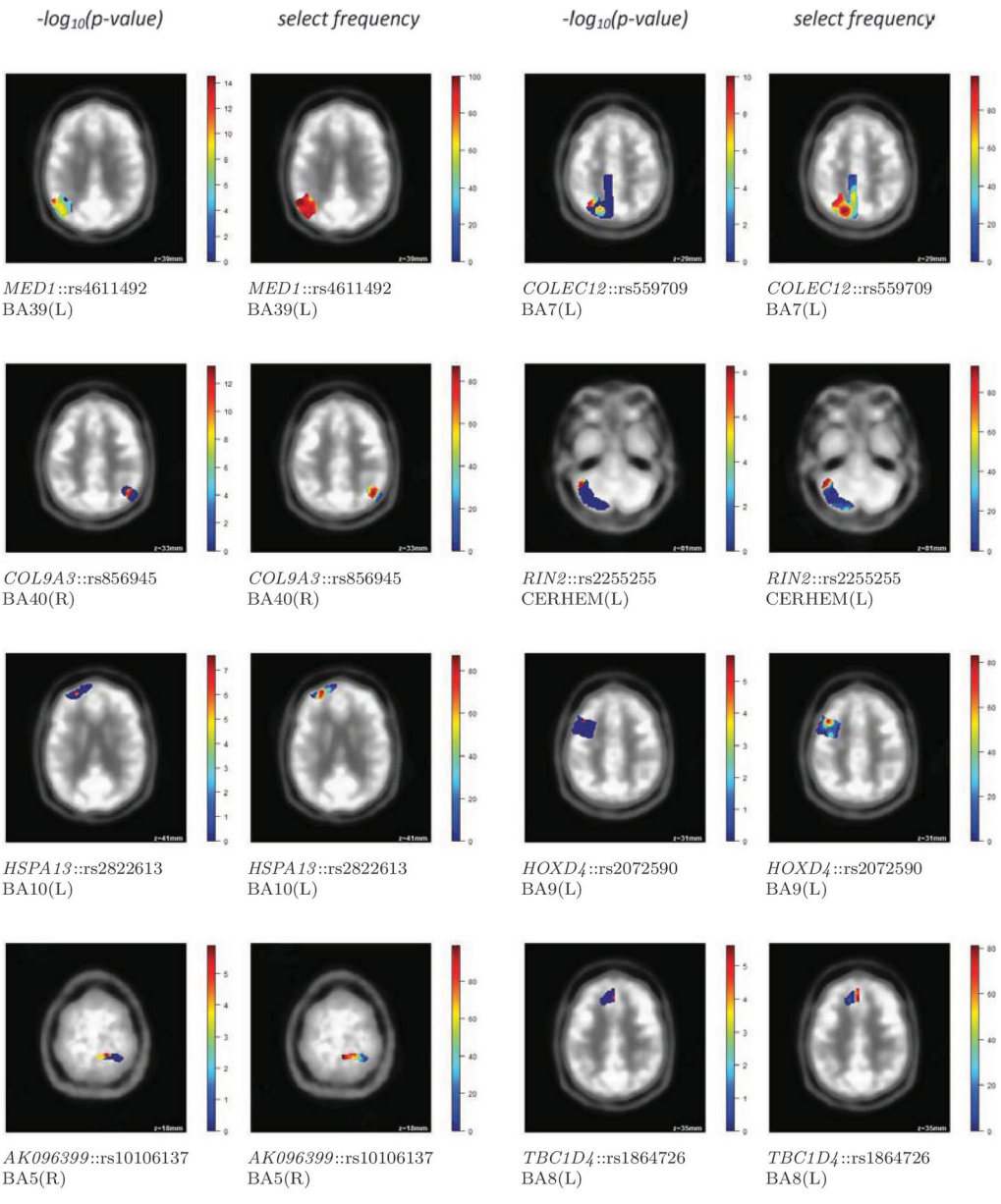
PRIMA1::rs4900194
BA38(R)

PRIMA1::rs4900194
BA38(R)

SNX29::rs10500383
BA21(R)

SNX29::rs10500383
BA21(R)

cjs_11605_figure2i.eps



cjs_11605_figure2ii.eps