Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2019 ; 16(6): 1986–1996. doi:10.1109/TCBB.2018.2833487.

Identifying Candidate Genetic Associations with MRI-Derived AD-Related ROI via Tree-Guided Sparse Learning

Xiaoke Hao,
School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

Xiaohui Yao,
Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104.

Shannon L. Risacher,
Department of Radiology and Imaging Sciences, School of Medicine, Indiana University, Indianapolis, IN 46202.

Andrew J. Saykin,
Department of Radiology and Imaging Sciences, School of Medicine, Indiana University, Indianapolis, IN 46202.

Jintai Yu,
Department of Neurology, Qingdao Municipal Hospital, School of Medicine, Qingdao University, Qingdao, Shandong 266000, China.

Huifu Wang,
Department of Neurology, Qingdao Municipal Hospital, School of Medicine, Qingdao University, Qingdao, Shandong 266000, China.

Lan Tan,
Department of Neurology, Qingdao Municipal Hospital, School of Medicine, Qingdao University, Qingdao, Shandong 266000, China.

Li Shen,
Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104.

Daoqiang Zhang
School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

Abstract

Imaging genetics has attracted significant interests in recent studies. Traditional work has focused on mass-univariate statistical approaches that identify important single nucleotide polymorphisms

(Corresponding authors: Li Shen and Daoqiang Zhang.), Li.Shen@pennmedicine.upenn.edu, dqzhang@nuaa.edu.cn.

(SNPs) associated with quantitative traits (QTs) of brain structure or function. More recently, to address the problem of multiple comparison and weak detection, multivariate analysis methods such as the least absolute shrinkage and selection operator (Lasso) are often used to select the most relevant SNPs associated with QTs. However, one problem of Lasso, as well as many other feature selection methods for imaging genetics, is that some useful prior information, e.g., the hierarchical structure among SNPs, are rarely used for designing a more powerful model. In this paper, we propose to identify the associations between candidate genetic features (i.e., SNPs) and magnetic resonance imaging (MRI)-derived measures using a tree-guided sparse learning (TGSL) method. The advantage of our method is that it explicitly models the complex hierarchical structure among the SNPs in the objective function for feature selection. Specifically, motivated by the biological knowledge, the hierarchical structures involving gene groups and linkage disequilibrium (LD) blocks as well as individual SNPs are imposed as a tree-guided regularization term in our TGSL model. Experimental studies on simulation data and the Alzheimer's Disease Neuroimaging Initiative (ADNI) data show that our method not only achieves better predictions than competing methods on the MRI-derived measures of AD-related region of interests (ROIs) (i.e., hippocampus, parahippocampal gyrus, and precuneus), but also identifies sparse SNP patterns at the block level to better guide the biological interpretation.

Keywords

Imaging genetics; tree-guided sparse learning; SNPs; hierarchical structure; ROIs

1 Introduction

NEUROIMAGING genetics is an emergent research field aiming at identifying genetic variants that influence measures derived from anatomical or functional brain images [1]. Compared to diagnostic measures based on cognitive or clinical assessments [2], [3], any quantitative measures extracted from different brain imaging modalities can be treated as intermediate or endophenotypes that are closer to the underlying biological mechanisms of the disease.

Genome-wide association studies (GWAS) in imaging genetics domain are increasingly being used to identify the associations between the high-throughput single nucleotide polymorphisms (SNPs) and the quantitative traits (QTs) of imaging data [4], [5]. To our knowledge, pairwise univariate analysis methods that work on statistic tests or p-values from standard individual SNP tests focus on examining statistical effects of each individual genetic variant at one time. However, it may ignore the underlying interacting relationship among SNPs and thus easily lead to a weak detection of associations. To address this problem, multivariate or multi-locus methods (e.g., by incorporating multi-SNP dependencies in the model) can detect SNPs missed by univariate methods [6], [7].

To jointly evaluate multiple correlated SNPs, regularization techniques such as ridge regression have been adopted in [8]. Meanwhile, some sparsity-based feature selection methods such as the L1-regularized least absolute shrinkage and selection operator (Lasso) [9] have been proposed to identify a subset of features (i.e., SNPs) for subsequent

association analysis. In [10], Lasso regression has been used to evaluate gene effects in a GWAS of the temporal lobe volume and discover a small set of genes that pass the genome-wide significance. In addition, Elastic Net, which combines L1-norm and L2-norm regularization to address the problem of high dimensionality and multiple colinearity simultaneously, has been implemented for imaging genetics studies [11], [12]. In particular, the L1-regularized term is used to enforce the 'sparsity' on the individual features, ignoring the structure information among SNPs that exist throughout the whole genome. Recently, based on Group Lasso by imposing the 'group sparsity' with L1/L2-norm regularization [13], an alternative method has been proposed in [14] to consider the group structure among SNPs. These L1 or L1/L2-regularized regression methods allow for a large number of correlated SNPs being incorporated into a single model and select a sparse set of SNPs that are associated with imaging measures. However, in the above methods, the hierarchical structure among SNPs that is different from flat group structure, is still not used for designing more powerful model.

On the other hand, in machine learning community, sparse learning methods with tree-structured regularizations have been proposed to model the underlying multilevel tree (i.e., hierarchical) structures among the inputs or outputs [15], [16]. The hierarchy- structured sparsity has been implemented with hierarchical agglomerative clustering technique for multi-scale mining on fMRI application [17]. Recently, the tree structure-based method has also been successfully used for neuroimaging-based brain disease classification [18].

Motivated by the above literature review, following the existing work (i.e., [19], [20]), we propose to identify more significant and meaningful SNP associated with magnetic resonance imaging (MRI)-derived measures from preselected candidate SNP set by using a tree-guided sparse learning (TGSL) method, which explicitly models the prior hierarchical tree structure among the SNPs in the objective function for feature selection. Here, the hierarchical tree structure is constructed based on the following prior knowledge, i.e., each tree node is for one feature group and different tree heights represent different levels of groups. Specifically, some SNPs are naturally connected via different pathways, and multiple SNPs located in one gene often jointly express certain genetic functionalities. On the other hand, another genetic biology phenomenon, i.e., linkage disequilibrium (LD) [21], describes the non-random distributions between alleles at different loci. Inspired by the above prior knowledge, the spatial gene and LD relationships among SNPs can be encoded into the tree regularization simultaneously that is an enhanced model for extending the previous work [22] to guide the feature selection for subsequent prediction. We demonstrate the practical utility of our method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort for identifying genetic associations with QTs of MRI-derived ROI measures (including bilateral volumes of the hippocampus, parahippocampal gyrus and precuneus) that are known to be relevant to Alzheimer's disease. From empirical experiment results, as expected, the tree-guided sparse learning not only yields improved prediction performances and but also identifies high-level SNP clusters jointly affecting relevant QTs.

It is worth noting that the focus of this initial study is to examine the prediction and feature selection power of the proposed TGSL model using a candidate SNP set, which performs a "pre-selection" step to get a biologically relevant candidate set with a moderate number of

SNPs. In the existing imaging genetic studies, there are many other works considering the entire brain regions or more interestingly interaction between image phenotype and the genotype simultaneously. Silver et al. [23] and Zhu et al. [24] proposed structured sparse low-rank regression models for imaging genetics analysis. Batmanghelich et al. [25], [26] and Zhu et al. [27] proposed Bayesian frameworks to identify multiple imaging phenotypes related to genetic markers. Both methods share a similar goal to identify relevant imaging phenotypes via multivariate multiple regression, which is different from our model where only one targeted phenotype associated to structured genotypes is analyzed in this study.

2 Method

2.1 Background of Sparse Learning on Imaging Genetics

Assume that there are M training subjects, with each represented by an N-dimensional feature vector (i.e., SNPs) and a response value (i.e., MRI-derived measure). Let X be an $M \times N$ feature matrix with the m-th row $x^m = (x_1^m, \ldots, x_n^m, \ldots, x_N^m) \in R^N$ denoting the m-th subject's feature vector, and y be the corresponding MRI-derived measures of M subjects. A linear regression model can be formulated as follows:

$$\mathrm{y} = \boldsymbol{X}\boldsymbol{\alpha} + \varepsilon, \tag{1}$$

where α is a vector of coefficients assigned to the respective features, and ε is an error term. To encourage the 'sparsity' among features, the L1-norm regularization is imposed on the vector of regression coefficients as follows [9]:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|y - \boldsymbol{X}\boldsymbol{\alpha}\|^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{2}$$

where λ is a regularization parameter that controls the sparsity in the solution. The non-zero elements in α indicate that the corresponding input features are relevant to the regression outputs. This L1-regularized Lasso regression method imposes sparisty on the individual variables for feature selection, which provides an effective multiple regression model to identify a subset of relative SNPs associated with MRI-derived measures. However, the Lasso-based method completely ignores the joint associated features, since another hypothesis is that the group or block of SNPs can convey important biological information and jointly affect the phenotypes.

In order to address the group-wise association among the features, sparsity can be enforced at the group level by an L1/L2-regularization, where the L2-norm is applied for the input features within the same group, while the L1-norm penalty is applied over the groups of input features [13], and it can be formulated as follows:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|y - \boldsymbol{X}\boldsymbol{\alpha}\|^2 + \lambda \sum_{j=1}^{N} w_j \|\boldsymbol{\alpha}_{G_j}\|_2, \tag{3}$$

where G_j (j = 1, …, N) is a set of pre-defined non-overlapping feature clustering groups, w_j is a predefined weight for the corresponding group G_j. The regularization term $\sum_{j=1}^{N} w_j \|\boldsymbol{\alpha}_{G_j}\|_2$, which refers to L1-norm on the $\|\alpha_{G_j}\|_2$ penalizes all coefficients in the

same group for joint feature selection. In practice, structure relationships are available when the priori knowledge are embedded. The group sparsity technique has been used in prior imaging studies [14]. Nevertheless, the disadvantage of this model is that it imposes the sparsity on the group-level while we know that perhaps only a handful of SNPs in a group are related.

Considering the limitations of L1-regularized Lasso and L1/L2-regularized Group Lasso, how can we identify the sparse solutions from a group with highly correlated variables? A recent regularization regression in reverse inference named as Elastic Net, has been proposed to conduct the high dimensional and correlated data analysis [11], [12] as following:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|y - \boldsymbol{X}\boldsymbol{\alpha}\|^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 + \lambda_2 \|\boldsymbol{\alpha}\|_2^2 . \tag{4}$$

The quadratic penalty term makes the loss function strictly convex, and therefore, it has a unique minimum. The Elastic Net method formulated as (4) combines the L1-regularized Lasso and ridge penalization. Each of them is a special case of the Elastic Net: (1) Lasso when $\lambda_1 = \lambda$ and $\lambda_2 = 0$; and (2) ridge when $\lambda_1 = 0$ and $\lambda_2 = \lambda$. In this context of imaging genetics, Elastic Net is indeed a competitive method to discover multi-locus associations.

2.2 Tree-Guided Sparse Learning

One limitation of the Lasso, Group Lasso and Elastic Net is that these methods have not taken into account the spatial structure of the data, and thus ignore the biological fact existing the hierarchal structure among SNPs. With these observations, in this section, we introduce a TGSL method [15] for solving the problem of identifying SNPs with hierarchical tree structures.

The hierarchical tree is constructed with intuitions that each tree node is for one feature group and different tree heights represent different levels of groups. The group construction is induced by the prior knowledge, i.e., multiple SNPs located in one gene often jointly affect certain genetic functionalities or alleles at different loci exhibit the non-random distributions with LD. Therefore, in the TGSL model, a tree structure is used to represent the hierarchical spatial relationship (grouping by LD blocks and by genes) among SNPs, with leaf nodes denoting SNPs and internal nodes denoting the groups of SNPs and groups of LD blocks. A schematic diagram in Fig. 1 shows an example of tree hierarchy structure.

Assume that a hierarchical tree T has d depth levels, and there are n_i nodes organized as $T_i = \{G_1^i, \ldots, G_j^i, \ldots, G_{n_i}^i\}$ in the ith level ($0 \le i \le d$). Different depth levels indicate the different scales of feature groups. The index sets of the nodes at the same level have no overlapping, and the index set of a child node is a subset of its parent's index set. The TGSL method [15] can be formulated as:

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|y - X\boldsymbol{\alpha}\|^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i \|\boldsymbol{\alpha}_{G_j^i}\|_2 . \tag{5}$$

where $\boldsymbol{\alpha}_{G_j^i}$ is the set of coefficients assigned to the features within node G_j^i, w_j^i is a predefined weight for node G_j^i with priori knowledge. Since each node presents a subtree of T, if one node is selected, all its descendant child nodes in tree will also be selected. A regularization predefined by the tree structure can be imposed on the sparse learning optimization problem to encourage a joint selection of structured relevant SNPs. In addition, the TGSL method combines the L1-regularized Lasso defined as (2) and non-overlapped Group Lasso defined as (3). Note that, it includes our previous formulation (3) as a special case, when the index tree is of depth 1 and $w_1^0 = 0$.

2.3 An Analytical Solution and Optimization

The objective function (5) can be efficiently solved using the Nesterov's accelerated proximal gradient optimization algorithm [28], [29], where the regularization also needs to be evaluated in each of its iteration. A detailed description on Moreau-Yosida regularization for grouped tree sparse learning optimization used in this paper can be found in [15].

The objective function can be separated into a smooth part and a non-smooth part. Accordingly, to achieve the approximating operation, the following function is constructed for approximating the composite function [28]:

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) + \langle \boldsymbol{\alpha} - \boldsymbol{\alpha}_i, \nabla f(\boldsymbol{\alpha}_i) \rangle + \frac{l}{2} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_i\|^2 + \lambda \Omega(\boldsymbol{\alpha}) . \tag{6}$$

The quadratic term keeps the update in a neighborhood, where f is close to its linear approximation. $\nabla f(\alpha_i)$ denotes the gradient of $f(\alpha)$ on point α_i at the ith iteration, and l is the step size which is an upper bound on the Lipschitz constant of ∇f. This problem can be equivalently re-written as:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{\alpha} - \left(\boldsymbol{\alpha}_i - \frac{1}{l} \nabla f(\boldsymbol{\alpha}_i)\right)\|^2 + \lambda \Omega(\boldsymbol{\alpha}) . \tag{7}$$

The proximal operator associated with our regularization term $\lambda\Omega$ is the function that maps a vector u to the unique solution. The grouped tree structure regularization described in [15] for a given v is formulated by:

$$\min_{u} \frac{1}{2} \|u - v\|^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i \|\boldsymbol{\alpha}_{G_j^i}\|_2 . \tag{8}$$

This operator is initially introduced to generalize the projection operator onto a convex set. What makes proximal methods appealing for solving sparse decomposition problems is that this operator can be computed in closed-form. Note the minimizer of (8) has the same solution as the original problem of (5). To implement the algorithm, we set $u^{d+1} = v$, and only need to maintain a working variable u, which is initialized with v for the minimization of (8) admits an analytical solution. We then traverse the index tree T in the reverse breadth-

first order to update u. At the traversed node G_j^i, we update u_G according to the operation in (10), with i from d to 0 and j from 1 to n_i. In addition, the implementation of the algorithm can also help explain why the structured group sparsity can be induced. Algorithm 1 shows such an efficient solution.

2.4 Imaging Phenotype Prediction

We treat each SNP as a feature and each QT as a response variable. And the similar regression model including multiple features (SNPs) and single response (imaging-derived measures) is also formulated in [30]. Based on the selected SNPs using TGSL, a regression model can be used for the final prediction on the phenotype. Support vector regression (SVR) is one of the widely used regression models to achieve generalized performance as it attempts to minimize the generalized error bound that is the combination of the training error and a regularization term with controlling the complexity of the hypothesis space [31]. Therefore, a linear kernel, which maps from original space to feature space is adopted for simplicity. The identification of QTs from genotype to phenotype is important to understand the underlying biological mechanism for the disease. In addition, our goal is to reveal the relationship between these identified SNP loci and imaging phenotypes. Thus, the genetic biomarkers we find in the prediction model with feature selection can provide therapists with a powerful and supplementary epidemiological evidence to assess the predisposition for a population to a certain disease (i.e., Alzheimer's disease (AD)).

Algorithm 1. To Minimize J in Equation (5)

Input: $v \in R^p$, the index tree T with nodes G_j^i, $\lambda > 0, \lambda_j^i = \lambda w_j^i, w_j^i \geq 0, (i = 0, 1, \ldots, d, j = 1, 2, \ldots, n_i)$

Output: $u^0 \in R^p$

1: Set:

$$u^{d+1} = v \tag{9}$$

2: **For** $i = d$ to 0

3: **For** $j = 1$ to n_i

4: Compute

$$u_{G_j^i}^i = \begin{cases} 0 & \left\| u_{G_j^i}^{i+1} \right\|_2 \leq \lambda w_j^i \\ \dfrac{\left\| u_{G_j^i}^{i+1} \right\|_2 - \lambda w_j^i}{\left\| u_{G_j^i}^{i+1} \right\|_2} u_{G_j^i}^{i+1} & \left\| u_{G_j^i}^{i+1} \right\|_2 > \lambda w_j^i \end{cases} \tag{10}$$

5: **End**

6: **End**

Fig. 2 shows the flowchart of the proposed method. Firstly, to capture the hierarchical relationship of the SNPs in our candidate set, we construct a tree structure by consisting of multiple levels. (1) The raw genetic data features (i.e., SNPs) are at the first (lowest) level. (2) The LD blocks are at the second level, where each LD block contains a set of SNPs. (3)

Genes are at the third level. Note that the SNPs studied in this work are extracted from risk genes. Thus, each gene contains a set of LD blocks and each LD block contains a set of SNPs. This forms a hierarchical structure among SNPs. Then, the constructed tree structure is imposed on the regularization of TGSL model to select the relevant features. Finally, SVR is used to predict the image-derived measures using the selected SNPs features.

3 Data Descriptions

In this section, we evaluate the effectiveness of the proposed method on the ADNI database (http://adni.loni.usc.edu/), where candidate SNPs are examined and selected to predict the response of the MR imaging phenotypes. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

In the present study, a total of 910 non-Hispanic Caucasian participants with both imaging and genotyping data are available, including 210 healthy control (HC), 82 significant memory concern (SMC), 272 early mild cognitive impairment (EMCI), 186 late mild cognitive impairment (LMCI), and 160 AD participants, as shown in Table 1.

3.1 Imaging Data and Pre-Processing

The MRI data used in this paper were obtained from the ADNI database. We aligned the preprocessed imaging data ((i.e., voxel based morphometry (VBM))) to each participant's same visit scan, and then created normalized gray matter density maps from MRI data in the standard Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm 3 voxels, registered by SPM software package [32]. 116 ROI level measurements of mean gray matter densities were further extracted based on the MarsBaR AAL atlas [33]. After removal of the cerebellum, the imaging measures of 90 ROIs were used as QTs in our experiments. All the measures were pre-adjusted for age, gender, education and intracranial volume (ICV) using the regression weights derived from the healthy control participants. We computed the volume of GM tissue in that ROI region as a feature.

Although brain images in different modality could provide a large number of different phenotypes for imaging genetic studies, selected candidate imaging traits should be highly heritable and be widely related with the pathology disease or biological process as closely as possible. Many image-derived measures such as bilateral volumes of the hippocampus, parahippocampal gyrus and precuneus are highly heritable and may be more directly influenced by genetic variation [34], [35], [36], [37]. The properties of the MRI-derived measurement responses are shown in Table 2.

3.2 Genotyping and Pre-Processing

Genotypes for the 910 subjects in this study were obtained from the ADNI database. The samples contain candidate genes from ADNI subjects, while genes information used for LD

calculation are from (http://browser.1000genomes.org/Homo_sapiens/UserData/Haploview?db=core).

Genome-wide genotyping data were available for the full set of ADNI subjects. All SNPs data, used in this study, are genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) [38]. In order to handle the large scale dirty genetic data, we used a well-known genetic analysis tool PLINK [39] to filter the genotype data using the following exclusion criteria: rare SNPs (minor allele frequency (MAF) < 0.05), violations of Hardy-Weinberg Equilibrium (HWE $p < 10^{-6}$), poor call rate (< 90%) per subject and per SNP marker, gender check, and sibling pair identification. Data were further "phased" to impute any missing individual genotypes after filtering using the MaCH program [40].

Following the approach to pre-selecting the SNPs, we used ANNOVAR(http://annovar.openbioinformatics.org) to annotate the SNPs with their corresponding genes listed in reference [19], [41] and the AlzGene database (http://www.alzgene.org). We focused our analysis on 20 AD risk genes. For each gene, we extracted all the SNPs within ±5k base pairs of the gene boundary based on the ANNOVAR annotation. Accordingly, there is no overlap among different groups.

This resulted in 3,781 SNPs being mapped to the top risk 20 genes. Fig. 3 presents the AD risk factor gene (*CR1, BIN1, INPP5D, MEF2C, EPHA1, NME8, ZCWPW1, CLU, PTK2B, CELF2, MS4A6A, SORL1, FERMT2, RIN3, SLC24A4, DSG2, ABCA7, CD33, APOE, CASS4*) and the numbers of preselected SNPs in our study.

As mentioned before in Fig. 1, we formed two interval levels in the tree. For the high interval nodes, since all SNPs had been divided into different genes naturally, we used the natural groups to construct groups, one for each of 20 genes.

In addition, for the low interval nodes, LD is another genetic biology phenomenon to construct groups, which refers to the non-random association of alleles at two or more loci. It is due to the physical connection between nearby loci on a chromosome, which is known as genetic linkage. Taking the SNPs on *APOE* for example, numerical values r2 of the LD maps were determined by Haploview [21], where r2 were the pairwise correlation between two SNPs as shown in Fig. 4. And Fig. 5 demonstrates the hierarchical structure grouping by LD blocks on APOE. Accordingly, on the LD levels, there were 233 blocks comprising 2,407 SNPs, with each of the remaining 1,374 SNPs being isolated by itself. For the input in the models, each SNP value was coded in an additive fashion as 0, 1 or 2, indicating the number of minor alleles.

4 Experimental Results

4.1 Simulation Study

In this section, we present a simulation study to show the potential power of the proposed TGSL model. We simulated data from the true model $y = X\alpha + \sigma\varepsilon$, where ε ($\varepsilon \sim N(0, 1)$) was the noise and σ was noise level (e.g., $\sigma = 0.01$). In this example, we set n = 100, p = 1,000, so that n < < p and we generated the n × p design matrix X from normal distribution

N(0, 1). And then we generated the vector α with a groupe structure (including 4 groups with all-zeros and 4 groups that were assigned to have sparse predictors) as follows:

$$\boldsymbol{\alpha}^{\mathrm{T}} = (\underbrace{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{20}}_{20}, \underbrace{0, \ldots, 0}_{180}, \underbrace{\boldsymbol{\alpha}_{21}, \ldots, \boldsymbol{\alpha}}_{10}, \underbrace{0, \ldots, 0}_{290},$$
$$\underbrace{\boldsymbol{\alpha}_{31}, \ldots, \boldsymbol{\alpha}_{40}}_{10}, \underbrace{0, \ldots, 0}_{290}, \underbrace{\boldsymbol{\alpha}_{41}, \ldots, \boldsymbol{\alpha}_{45}}_{5}, \underbrace{0, \ldots, 0,}_{195}).$$

Thus, we obtained the response y. In this experiment, we performed 3 sets of simulations with different sparsity levels, which respectively included 5, 15 and 25 truth signals out of 45 variables (i.e., those from the non-zeros groups).

In each study comparison, we used L1-regularized Lasso, L1/L2-regularized Group Lasso, Elastic Net and our proposed TGSL methods to select a subset of features and predict the regression responses, respectively. In simulations, we set the number of depth levels d in the tree structure as 2. Following the previous studies [15, 18], we set the weight assignments on the penalized groups as the square root of the group elements for the Group Lasso and the proposed TGSL model. All regularization parameters from models (including Lasso, Group Lasso, Elastic Net and TGSL) were optimally tuned using a grid search from the range of {0, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1} by nested 5-fold cross-validation on the training set. The performances on each simulation dataset were assessed with root mean squared error (RMSE), Pearson correlation coefficient (PCC) and coefficient of determination (CD). In our experiments, 5-fold cross-validation strategy was adopted to evaluate the effectiveness of the reference methods. For facilitating efforts to replicate our results, we have published the matlab code and simulation data. These resources are available at https://sourceforge.net/projects/ibrain-cn/files/TGSL.

As shown in Table 3, the TGSL for all signal amounts consistently outperforms the other methods in RMSE measurement. Regarding the PCC results in Table 4, the best PCCs are obtained by TGSL and it shows the best PCCs of 0.8 for all signal sparsity levels. It is worth noting that without an explicit apriori knowledge, Elastic Net can get better performances than Lasso, since the L2 penalty shrinks the coefficients particularly in highly correlated features and thus indirectly encourages the grouping effect [11], [12]. While compared to Group Lasso, TGSL can well impose the sparsity on the group levels. Here, we are aware of the fact supporting the scene that only a handful of features in a group are related. As shown in Table 5, although TGSL shows the weak predictabilities, it is better than the other methods in CD measurement. These experiment results quantified by different measurements demonstrate that the proposed tree-guided structure can help improve regression performance and discover the signals better than the other methods.

4.2 Real Data Applications on ADNI

In the real data applications, we used L1-regularized Lasso, L1/L2-regularized Group Lasso, Elastic Net and our proposed TGSL methods to select a subset of features (i.e., SNPs) to predict the regression responses on the test data. In this experiment setting, we set the number of depth levels d in the tree structure as 3. As for the Group Lasso and our proposed TGSL, we set the weight assignments on the penalized groups as the square root of the

group elements' size. Since L1/L2-regularized method refers to the flat manner, we defined the groups using LD blocks for Group Lasso.

The primary goal of this experiment is trying to demonstrate the superiority of our proposed TGSL when all methods select the same number of features. Although the Lasso and Elastic Net are considered as flexible dimensionality reduction algorithms to select any number of features (e.g., the fewest or only one) for different sparsity levels, from the perspective of system biology, it is more meaningful to discover a set number of loci as well as genes rather than top significant SNPs for pathway analysis and interpretations. On the other hand, as the hierarchical clustering structures are imposed for joint effectiveness consideration in optimization, both Group Lasso and our proposed TGSL methods won't be able to select fewer features than the size of the smaller feature group. Thus, for fair comparisons, we have only present the performances of the compared methods in terms of their capability to select the same numbers of features in the moderate areas. As for the regularized parameters in all comparisons, we determined their values corresponding to the number of selected SNPs from 200 to 2000 with the approximate step of hundreds. The performance of each trial was assessed with RMSE, a widely used criterion in regression analysis. Average RMSE result was calculated based on 10-fold cross validation.

4.2.1 Regression Results Using Selected SNPs—We compare our proposed TGSL methods with standard feature selection methods such as Lasso, group Lasso and Elastic Net. For testing the regression performance with respect to different level of selected features in all methods, we adjust the regularization parameter to control the sparsity. Fig. 6 reports the RMSE for regression on the bilateral (i.e., left and right) volumes of hippocampaus, parahippcampal gyrus and precuneus, respectively, by adopting a polynomial model to fit all the data obtained with different regularized parameters. According to Fig. 6, the proposed TGSL methods outperform the other competing methods on all the tested ROIs, showing the promise of TGSL. It's worth noting that RMSE increased steeply as a function of the number of selected SNPs because the extra SNPs are noisy ones and thus harmful for the regression performance. Our results are in accordance with that of a prior study [19].

4.2.2 SNP Biomarker Selections—Following the previous regression results on the volumes of brain ROIs calculated on average, taking left hippocampus and right parahippocampal gyrus for example, we selected the 200 SNPs associated with the MRI-derived measures from one fold test in one trial. As shown in Table 6, the proposed TGSL can achieve the best RMSE value on MIR-derived ROI volume predictions comparing to the other competitive methods.

As illustrated in Fig. 7, some relevant SNPs detected by Lasso method are observed on widespread gene space. There are few robust SNPs in feature selection via Group Lasso. It imposes the sparsity on the group-level while we know that only a handful of SNPs in a group are related. Elastic Net can identify the sparse solutions from a group with highly correlated SNPs, comparing to L1-regularized Lasso.

We can also observe from the figure that as inducing the tree structure, the TGSL solutions lead to sparse blocks, where the selected SNPs are from meaningful LD blocks and two risk genes (*INPP5D* and *CD33*). As expected, the tree-guided sparse learning is to select multiple SNPs from one gene to find the joint effectiveness on synthesis of functional proteins in metabolic process. For example, *INPP5D* has been reported to be associated with AD through modulating the inflammatory process and immune response [42]. As a type I transmembrane protein, *CD33* belongs to the sialic acid-binding immunoglobulin-like lectins, mediating the cell–cell interaction and inhibiting normal functions of immune cells. In the brain, it is mainly expressed on microglial cells. The level of *CD33* has been found to be increased in the AD brain, which positively correlates with amyloid plaque burden and disease severity [43]. It is worth noting that these stable markers are also reported in other related heritable neurodevelopmental disorders, which is in accordance with the existing findings [44]. Although they have been considered as risk factor genes in dementia, the imaging genetic finding of joint *INPP5D* and *CD33* warrants further investigation.

5 Conclusion

In this paper, we investigate the potential of exploiting tree-guided sparse learning (TGSL) method for identifying the associations between candidate SNPs and AD-related MRI-derived ROI (hippocampus, parahippocampal gyrus and precuneus) measures, given hierarchical tree structure information (i.e., gene groups and LD blocks as well as individual SNPs). The experimental results on simulation studies and real data ADNI applications show the better performance that demonstrates the tree-guided regularizations help to discover the marks better than other reference methods. Furthermore, the similar model can be extended and applied to the other MRI-derived measure ROIs or other additional modality phenotypes (e.g., DTI, fMRI and PET data).

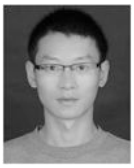
Since the TGSL study has focused on single phenotype outcome, it is an interesting future topic to expand this model into the multivariate multiple regression frameworks considering whole brain or more complex interaction between genotype, phenotype and diagnosis factors to identify more complicated multi-SNP-multi-phenotype associations. In addition, another interesting problem is If we extend upstream and downstream threshold (e.g., > 20 kbp) on the gene annotation, some SNPs could be annotated by multiple genes. While the non-overlapping group structure in group Lasso limits its applicability in practice. Thus, the overlapping group Lasso penalized problem [45] or graph structure associations [46] should be further investigated in our future work.

Acknowledgments

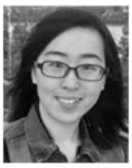
Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Bio-gen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda

Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research was supported by the NSFC (Nos. 61422204, 61473149, 61732006) and TSPP (No. 17ZLZDZF00040) in China. At Indiana University, this work was supported by NIH R01 EB022574, R01 LM011360, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG 042437, and R01 AG046171; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; NCAA 14132004; and CTSI SPARC Program.

Biography



Xiaoke Hao received the BS and MS degrees from Nanjing University of Information Science and Technology, Nanjing, China, in 2009 and 2012, respectively, and the PhD degree in computer science and technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2017. He is currently an assistant professor with the School of Artificial Intelligence of the Hebei University of Technology. His research interests include machine learning and imaging genetics. He is a member of the IEEE.



Xiaohui Yao received the BS degree in computer science and technology from Qing Dao University, and the MS degree in computer software and theory from the University of Science and Technology of China. She is currently working toward the PhD degree in bioinformatics at Indiana University-Purdue University Indianapolis, and is a postdoctoral fellow with the Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine School of Medicine. Her research interests include imaging genetics, multidimensional data mining, and information visualization.

Shannon L. Risacher received the BS degree in psychology from Indiana University-Purdue University Indianapolis, and the PhD degree in medical neuroscience from the Indiana University School of Medicine. She is an assistant professor of radiology and imaging sciences with the Indiana University School of Medicine. Her main research interests involve evaluating imaging and non-imaging biomarkers of Alzheimer's disease

(AD) for utility in early detection and diagnosis. In particular, she is interested in evaluating which biomarkers are most sensitive in the earliest stages of disease, both for detecting pathophysiological changes and for predicting future clinical outcomes. She is primarily focused on structural, functional, and molecular imaging biomarkers of AD, but has an additional interest in novel biomarkers such as sensory and perceptual tests.

Andrew J. Saykin received the BA degree in psychology from the University of Massachusetts Amherst, and the MS degree in clinical psychology, and the PsyD degree in clinical neuropsychology from Hahnemann Medical College. He is the Raymond C. Beeler professor of radiology and professor of medical and molecular genetics with the Indiana University School of Medicine. His expertise is in the areas of multimodal neuroimaging research, human genetics, and neuropsychology/cognitive neuroscience. He has a longstanding interest in the structural, functional, and molecular substrates of cognitive deficits in Alzheimer's disease, cancer, brain injury, schizophrenia, and other neurological and neuropsychiatric disorders. The major thrust of his current research program is on integrating advanced brain imaging and genomic data to enhance the understanding of disorders affecting memory.

Jintai Yu received the MD and MS degrees from the School of Medicine, Qingdao University, Qingdao, China, in 2006 and 2009, respectively, and the PhD degree from Nanjing Medical University, Nanjing, China, in 2015. He accomplished his postdoctoral research with the University of California, Los Angeles, in 2016. He is currently a professor with the Department of Neurology, Qingdao University, specializing in dementia. His current main research fields are: (1) to investigate the genetics, biomarkers, and neuroimaging of Alzheimer's disease: and (2) to explore the pathogenesis and novel therapy of Alzheimer's disease.

Huifu Wang received the MD degree in neurology from Qingdao Municipal Hospital, School of Medicine, Qingdao University, Qingdao, China, in 2010, and the PhD degree in neurology from Qingdao Municipal Hospital, Nanjing Medical University, Nanjing, China.

He is currently a resident with the Department of Neurology, Qingdao University, specializing in dementia. His current research interests mainly include imaging genetics in Alzheimer's disease.

Lan Tan received the graduated degree from the Department of Medicine in Qingdao Medical College, in 1983, and the PhD degree, in 2009. In her studying period, she has visited Tel Aviv University in Israel and King's College of London in United Kingdom as a senior visiting scholar. Currently, professor Lan Tan works as the vice president of Qingdao Municipal Hospital, supervisor of the Brain Center, and head of the Neurological Department. In addition, she is the National Committee of Chinese Neurological Association, vice chairman of Shandong Provincial Neurological Association, judge of the National Nature Science Foundation of China, editorial board member of 10 Chinese and foreign academic journals, including the *Journal of Alzheimer's Disease, Chinese Journal of Neurology,* etc. She is mainly focused on research in dementia and epilepsy.

Li Shen received the BS degree from Xi'an Jiao Tong University, the MS degree from Shanghai Jiao Tong University, and the PhD degree from Dartmouth College, all in computer science. He is a professor with the Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine School of Medicine. His research interests include medical image computing, bioinformatics, data mining, network science, systems biology, brain imaging genomics, and brain connectomics. He is a member of the IEEE.

Daoqiang Zhang received the BS and PhD degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a lecturer, in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published more than 150 scientific articles in refereed international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Medical Imaging,*

Neuroimage, Human Brain Mapping, Medical Image Analysis; and conference proceedings such as IJCAI, AAAI, NIPS, CVPR, and MICCAI, with more than 8,000 citations by Google Scholar. He was nominated for the National Excellent Doctoral Dissertation Award of China, in 2006, won the best paper award or best student award of several international conferences such as PRICAI'06, STMI'12, and BICS'16, etc. He has served as a program committee member for several international and native conferences such as IJCAI, AAAI, NIPS, SDM, PRICAI, and ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).

References

[1]. Ge T, Schumann G, and Feng J, "Imaging genetics—towards discovery neuroscience," Quantitative Biology, vol. 1, pp. 227–245, 2013.

[2]. Gottesman II and Gould TD, "The endophenotype concept in psychiatry: Etymology and strategic intentions," Am. J. Psychiatry, vol. 160, pp. 636–645, 2003. [PubMed: 12668349]

[3]. Glahn DC, Thompson PM, and Blangero J, "Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function," Human Brain Mapping, vol. 28, pp. 488–501, 6 2007. [PubMed: 17440953]

[4]. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD, Huentelman MJ, Craig DW, DeChairo BM, Potkin SG, Jack CR, Weiner MW, Saykin AJ, and Initia ADN, "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," Neuroimage, vol. 53, pp. 1051–1063, 11 15 2010. [PubMed: 20100581]

[5]. Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, DeChairo BM, Potkin SG, Weiner MW, Thompson PM, and Initia ADN, "Voxelwise genome-wide association study (vGWAS)," Neuroimage, vol. 53, pp. 1160–1174, 11 15, 2010. [PubMed: 20171287]

[6]. Ballard DH, Cho J, and Zhao HY, "Comparisons of multimarker association methods to detect association between a candidate region and disease," Genetic Epidemiology, vol. 34, pp. 201–212, 4 2010. [PubMed: 19810024]

[7]. Hibar DP, Kohannim O, Stein JL, Chiang M-C, and Thompson PM, "Multilocus genetic analysis of brain images," Frontiers Genetics, vol. 2, 2011, Art. no. 73.

[8]. Kohannim O, Hibar DP, Stein JL, Jahanshad N, Jack CR, Weiner MW, Toga AW, and Thompson PM, "Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression," in Proc. IEEE Int. Symp. Biomed. Imaging: From Nano Macro, 2011, pp. 1855–1859.

[9]. Tibshirani R, "Regression shrinkage and selection via the lasso: A retrospective," J. Royal Statistical Society Series B-Statistical Methodology, vol. 73, pp. 273–282, 2011.

[10]. Kohannim O, Hibar DP, Stein JL, Jahanshad N, Hua X, Rajagopalan P, Toga AW, Jack CR Jr., Weiner MW, de Zubicaray GI, McMahon KL, Hansell NK, Martin NG, Wright MJ, and Thompson PM, "Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression," Front Neurosci, vol. 6, 2012, Art. no. 115.

[11]. Kohannim O, Hibar D, Jahanshad N, Stein J, Hua X, Toga AW, Jack CR, Weiner MW, Thompson PM, and Initia ADN, "Predicting temporal lobe volume on mri from genotypes using L-1-L-2 regularized regression," in Proc. 9th IEEE Int. Symp. Biomed. Imaging, 2012, pp. 1160–1163.

[12]. Zou H and Hastie T, "Regularization and variable selection via the elastic net," J. Royal Statistical Soc. Series B-Statistical Methodology, vol. 67, pp. 301–320, 2005.

[13]. Yuan M and Lin Y, "Model selection and estimation in regression with grouped variables," J. Royal Statistical Soc. Series B-Statistical Methodology, vol. 68, pp. 49–67, 2006.

[14]. Wang H, Nie FP, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L, and Initiative ASDN, "Identifying quantitative trait loci via group-sparse multitask regression and feature

selection: An imaging genetics study of the ADNI cohort," Bioinf., vol. 28, pp. 229–237, 1 15, 2012.

[15]. Liu J and Ye J, "Moreau-Yosida regularization for grouped tree structure learning," in Proc. Advances Neural Inf. Process. Syst, 2010, pp. 1459–1467.

[16]. Kim S and Xing EP, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL Mapping," Ann. Appl. Statistics, vol. 6, pp. 1095–1117, 9 2012.

[17]. Jenatton R, Gramfort A, Michel V, Obozinski G, Eger E, Bach F, and Thirion B, "Multiscale mining of fMRI data with hierarchical structured sparsity," SIAM J. Imaging Sci, vol. 5, pp. 835–856, 2012.

[18]. Liu M, Zhang D, Yap PT, and Shen D, "Tree-guided sparse coding for brain disease classification," Medical Image Comput. Comput. Assisted Intervention, vol. 15, pp. 239–47, 2012.

[19]. Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, and Shen L, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort," Bioinf., vol. 28, pp. 229–237, 1 15, 2012.

[20]. Wang H, Nie F, Huang H, Yan J, Kim S, Nho K, Risacher SL, Saykin AJ, and Shen L, "From phenotype to genotype: An association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs," Bioinf., vol. 28, pp. i619–i625, 9 15, 2012.

[21]. Barrett JC, Fry B, Maller J, and Daly MJ, "Haploview: Analysis and visualization of LD and haplotype maps," Bioinf., vol. 21, pp. 263–265, 1 15, 2005.

[22]. Hao X, Yu J, and Zhang D, "Identifying genetic associations with MRI-derived measures via tree-guided sparse learning," Medical Image Comput. Comput. Assisted Intervention, vol. 17, pp. 757–64, 2014.

[23]. Silver M, Janousova E, Hua X, Thompson PM, and Montana G, "Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression," Neuroimage, vol. 63, pp. 1681–94, 11 15, 2012. [PubMed: 22982105]

[24]. Zhu X, Suk H-I, Huang H, and Shen D, "Structured sparse low-rank regression model for brain-wide and genome-wide associations," in Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervention, 2016, pp. 344–352.

[25]. Batmanghelich NK, Dalca AV, Sabuncu MR, and Polina G, "Joint modeling of imaging and genetics," Inf. Process. Medical Imaging, vol. 23, pp. 766–777, 2013.

[26]. Batmanghelich NK, Dalca A, Quon G, Sabuncu M, and Golland P, "Probabilistic modeling of imaging, genetics and diagnosis," IEEE Trans. Medical Imaging, vol. 35, no. 7, pp. 1765–1779, 7 2016. [PubMed: 26886973]

[27]. Zhu H, Khondker Z, Lu Z, and Ibrahim JG, "Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers," J. Amer. Statistical Assoc, vol. 109, pp. 977–990, 2014.

[28]. Nesterov Y, "Gradient methods for minimizing composite functions," Math. Program, vol. 140, pp. 125–161, 8 2013.

[29]. Beck A and Teboulle M, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imaging Sci, vol. 2, pp. 183–202, 2009.

[30]. Yip WK and Lange C, "Quantitative trait prediction based on genetic marker-array data, a simulation study," Bioinf., vol. 27, pp. 745–748, 3 15, 2011.

[31]. Basak D, Pal S, and Patranabis DC, "Support vector regression," Neural Inf. Process.-Lett. Rev, vol. 11, pp. 203–224, 2007.

[32]. Ashburner J and Friston K, "Voxel-based morphometry," Statistical Parametric Mapping: The Analysis Functional Brain Images, pp. 92–98, 2007.

[33]. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, and Joliot M, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," Neuroimage, vol. 15, pp. 273–89, 1 2002. [PubMed: 11771995]

[34]. Thompson PM, Martin NG, and Wright MJ, "Imaging genomics," Current Opinion Neurology, vol. 23, 2010, Art. no. 368.

[35]. Jack CR Jr., Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, and Kokmen E, "Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment," Neurology, vol. 52, pp. 1397–403, 4 22, 1999. [PubMed: 10227624]

[36]. Karas G, Scheltens P, Rombouts S, van Schijndel R, Klein M, Jones A, van der Flier W, Vrenken H, and Barkhof F, "Precuneus atrophy in early-onset Alzheimer's disease: A morphometric structural MRI study," Neuroradiology, vol. 49, pp. 967–976, 12 2007. [PubMed: 17955233]

[37]. Echavarri C, Aalten P, Uylings HBM, Jacobs HIL, Visser PJ, Gronenschild EHBM, Verhey FRJ, and Burgmans S, "Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease," Brain Structure Function, vol. 215, pp. 265–271, 1 2011. [PubMed: 20957494]

[38]. Saykin AJ, Shen L, Foroud TM, Potkin SG, Swaminathan S, Kim S, Risacher SL, Nho K, Huentelman MJ, Craig DW, Thompson PM, Stein JL, Moore JH, Farrer LA, Green RC, Bertram L, Jack CR, Weiner MW, and Initi ASDN, "Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans," Alzheimers Dementia, vol. 6, pp. 265–273, 5 2010.

[39]. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, and Daly MJ, "PLINK: A tool set for whole-genome association and population-based linkage analyses," Am. J. Human Genetics, vol. 81, pp. 559–575, 2007. [PubMed: 17701901]

[40]. Li Y, Willer CJ, Ding J, Scheet P, and Abecasis GR, "MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes," Genetic Epidemiology, vol. 34, pp. 816–834, 12 2010. [PubMed: 21058334]

[41]. Bertram L, McQueen MB, Mullin K, Blacker D, and Tanzi RE, "Systematic meta-analyses of Alzheimer Disease genetic association studies: The AlzGene database," Nature Genetics, vol. 39, pp. 17–23, 1 2007. [PubMed: 17192785]

[42]. Jing H, Zhu JX, Wang HF, Zhang W, Zheng ZJ, Kong LL, Tan CC, Wang ZX, and Tan L, "INPP5D rs35349669 polymorphism with late-onset Alzheimer's Disease: A replication study and meta-analysis," Oncotarget, vol. 7, pp. 69225–69230, 2016. [PubMed: 27750211]

[43]. Jiang T, Yu JT, Hu N, Tan MS, Zhu XC, and Tan L, "CD33 in Alzheimer's Disease," Molecular Neurobiology, vol. 49, 2014, Art. no. 529.

[44]. Raj T, Ryan KJ, Replogle JM, Chibnik LB, Rosenkrantz L, Tang A, Rothamel K, Stranger BE, Bennett DA, Evans DA, De Jager PL, and Bradshaw EM, "CD33: Increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility," Human Molecular Genetics, vol. 23, pp. 2729–2736, 5 15, 2014. [PubMed: 24381305]

[45]. Yuan L, Liu J, and Ye J, "Efficient methods for overlapping group lasso," in Proc. Advances Neural Inf. Process. Syst, 2011, pp. 352–360.

[46]. Mairal J, Jenatton R, Bach FR, and Obozinski GR, "Network flow algorithms for structured sparsity," in Proc. Advances Neural Inf. Process. Syst, 2010, pp. 1558–1566.

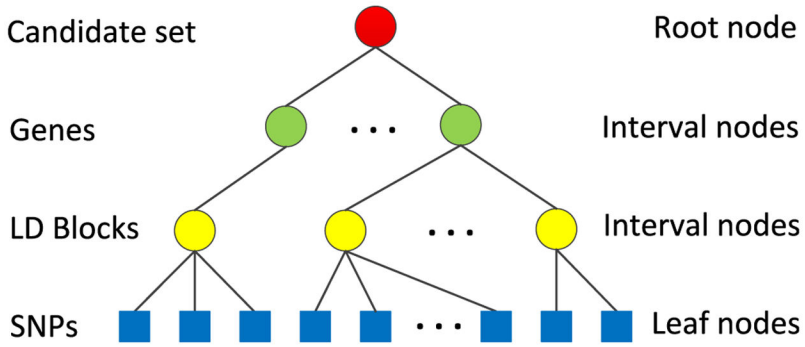


Fig. 1.

Illustration of the tree-structured hierarchical relationship among SNPs: group by gene and group by linkage disequilibrium (LD) blocks.

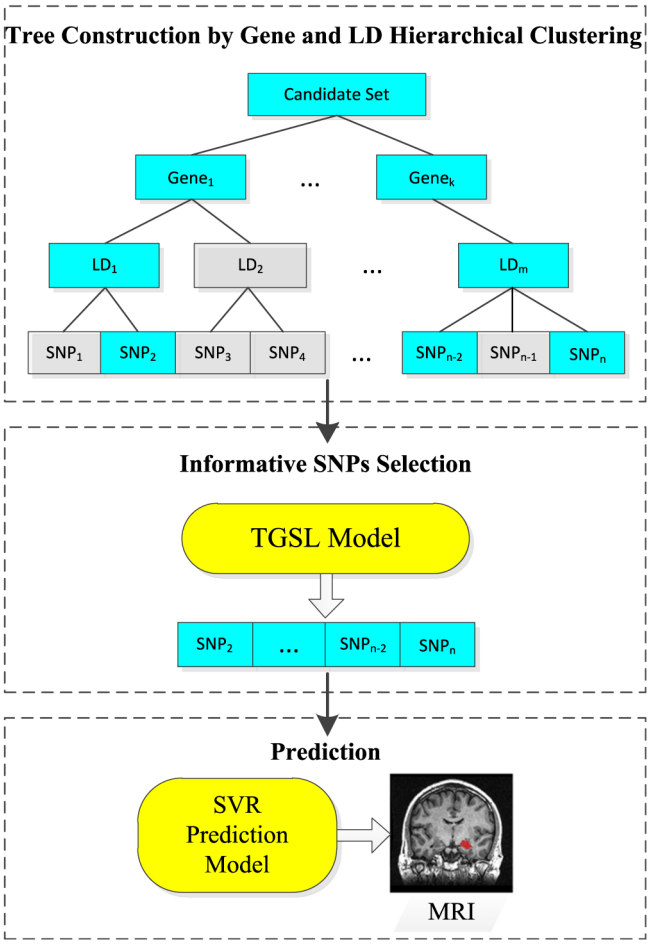


Fig. 2.
The flowchart of the proposed method.

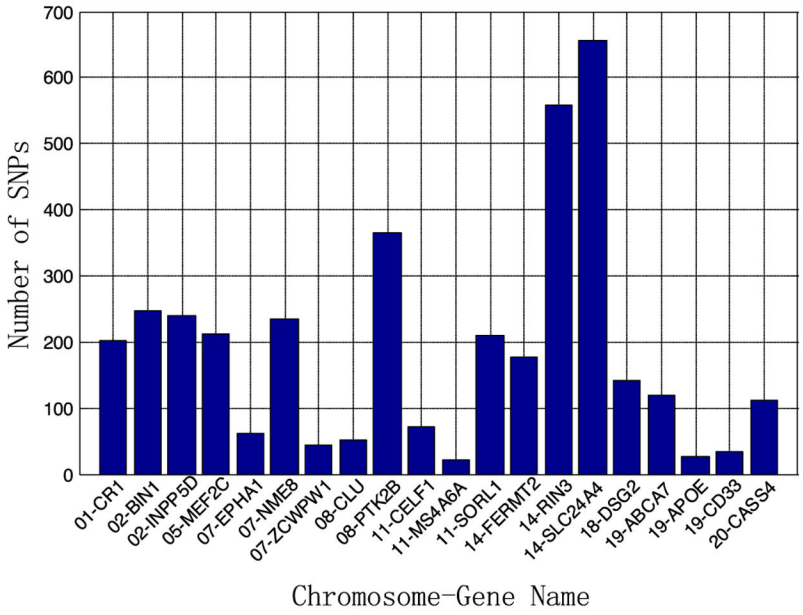


Fig. 3.
The Top 20 AD risk genes used in this study and the numbers of their SNPs.

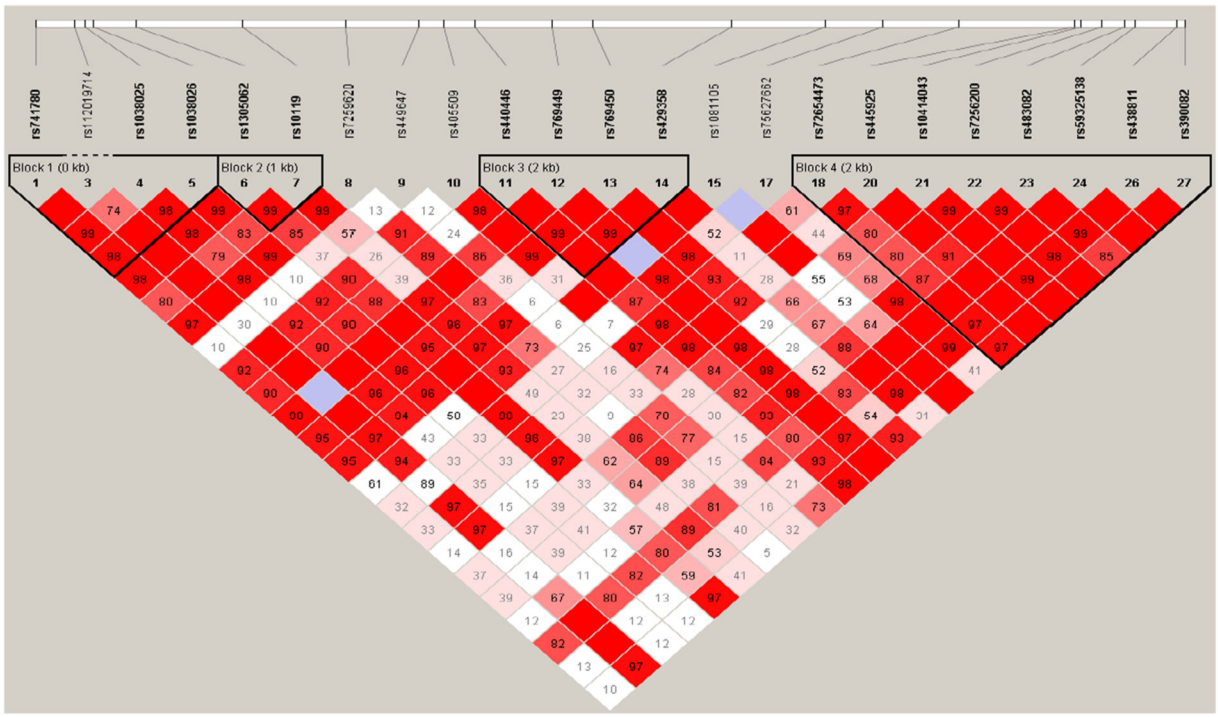


Fig. 4.
Illustration plot on LD of SNPs on *APOE* by Haploview.

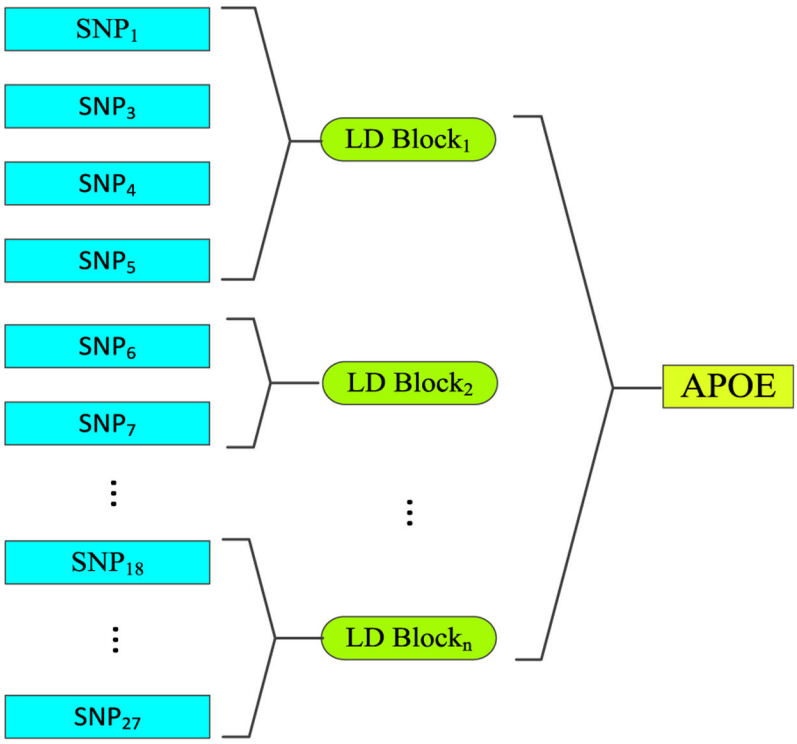


Fig. 5.
Hierarchical structure grouping by LD blocks on APOE.

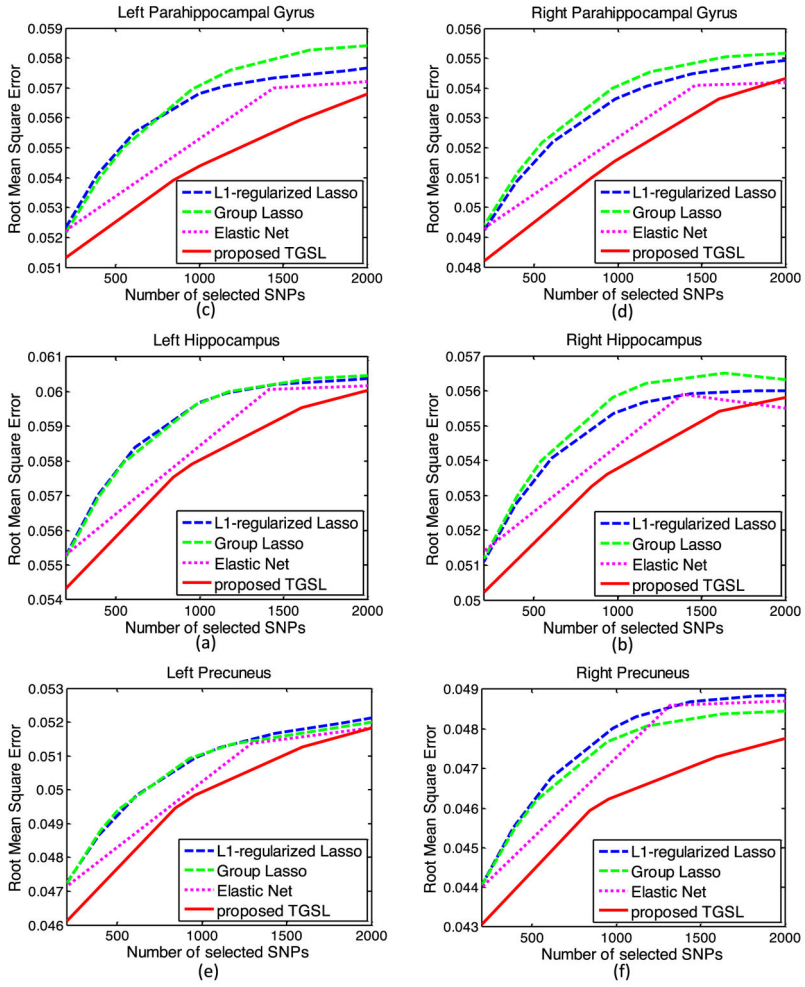


Fig. 6.

Comparison of RMSE with respect to different number of selected SNPs from 200 to 2000 by L1-regularized Lasso, Group Lasso, Elastic Net, the proposed TGSL in prediction on (a) Left Hippocampus, (b) Right Hippocampus, (c) Left Parahippocampal Gyrus, (d) Right Parahippocampal Gyrus, (e) Left Precuneus, and (f) Right Precuneus.

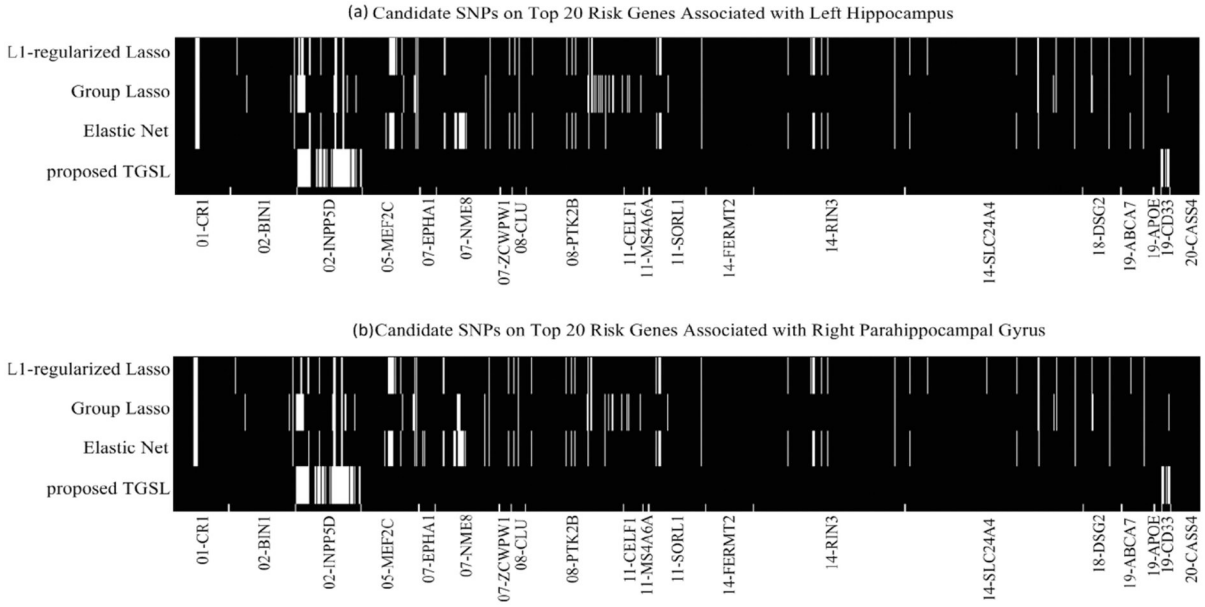


Fig. 7.

The patterns of SNP selections by L1-regularized Lasso, Group Lasso, Elastic Net, and the proposed TGSL on (a) Left Hippocampus and (b) Right Parahippocampal gyrus. The white entries are masked as selected SNPs.

TABLE 1

Characteristics of the Subjects

Subjects	HC	SMC	EMCI	LMCI	AD
Number	210	82	272	186	160
Gender	109	33	153	107	95
(M/F)	/101	/49	/119	/79	/65
Age	76.13	72.45	71.51	73.79	75.18
(mean±std)	±6.54	±5.67	±7.11	±8.40	±7.88
Education	16.44	16.78	16.07	16.38	15.86
(mean±std)	±2.62	±2.67	±2.62	±2.82	±2.75

Note: HC = Healthy Control, SMC = Significant Memory Concern, EMCI = Early Mild Cognitive Impairment, LMCI = Late Mild Cognitive Impairment, and AD = Alzheimer's disease.

TABLE 2

Properties of the MRI- Derived Measurement Responses

Response ROI	Range	Mean±Std
Left Hippocampus	0.268-0.598	0.473 ± 0.052
Right Hippocampus	0.245-0.567	0.441 ± 0.049
Left Parahippocampal Gyrus	0.303-0.686	0.509 ± 0.049
Right Parahippocampal Gyrus	0.350-0.712	0.554 ± 0.047
Left Precuneus	0.179-0.597	0.356 ± 0.045
Right Precuneus	0.178-0.598	0.371 ± 0.042

TABLE 3

Root Mean Squared Error on Simulation Study

#of True Signals	L1Lasso	Group Lasso	ElasticNet	proposed TGSL
5	0.70 ± 0.24	0.74 ± 0.19	0.69 ± 0.24	0.60 ± 0.22
15	3.66 ± 0.50	3.44 ± 0.61	3.58 ± 0.43	3.06 ± 0.60
25	4.85 ± 1.16	4.18 ± 1.20	4.81 ± 1.20	3.80 ± 0.91

TABLE 4

Pearson Correlation Coefficient on Simulation Study

#of True Signals	L1Lasso	Group Lasso	ElasticNet	proposed TGSL
5	0.81 ± 0.14	0.81 ± 0.12	0.82 ± 0.14	0.86 ± 0.14
15	0.41 ± 0.29	0.77 ± 0.07	0.48 ± 0.25	0.82 ± 0.08
25	0.14 ± 0.14	0.81 ± 0.09	0.20 ± 0.12	0.82 ± 0.12

TABLE 5

Coefficient of Determination on Simulation Study

#of True Signals	L1Lasso	Group Lasso	ElasticNet	proposed TGSL
5	0.61 ± 0.21	0.58 ± 0.16	0.62 ± 0.21	0.70 ± 0.22
15	0.04 ± 0.43	0.22 ± 0.08	0.07 ± 0.43	0.38 ± 0.05
25	−0.06 ± 0.06 *	0.22 ± 0.05	−0.04 ± 0.09 *	0.35 ± 0.07

Note:

* indicates a negative value (the sum of squares of residuals is larger than the variance of the data).

TABLE 6

RMSE with Respect to 200 Selected SNPs by Different Methods

Response ROI	Hippocampus_L	Parahipp_R
L1-Lasso	0.0555	0.0491
Group Lasso	0.0553	0.0501
Elastic Net	0.0557	0.0490
Proposed TGSL	0.0547	0.0487