An L_0 Regularization Method for Imaging Genetics and Whole Genome Association Analysis on Alzheimer's Disease

Xiong Li, Yangkai Lin, Xu Meng, Yangping Qiu, and Bo Hu

***Abstract*—Although the neuroimaging measures build a bridge between genetic variants and disease phenotypes, an assessment of single nucleotide variants changes in brain structure and their clinically influence on the progression of Alzheimer's disease remain largely preliminary. Note that each variant has very weak correlation signal to neuroimaging measures or Alzheimer's disease phenotypes. Therefore, traditional sparse regression-based image genetics approaches confront with unresolvable features, relative high regression error or inapplicability of high-dimensional data. Adopting an L_0 regularization method, we significantly elevate the regression accuracy of imaging genetics compared with group-sparse multitask regression method. With further analysis on the simulation results, we conclude that multiple regression tasks model may be unsuitable for image genetics. In addition, we carried out a whole genome association analysis between genetic variants (about 388 million loci) and phenotypes (cognition normal, mild cognitive impairment and Alzheimer's disease) with using the L_0 regularization method. After annotating the effect of all variants by Ensembl Variant Effect Predictor (VEP), our method locates 33 missense variants which can explain 40% phenotype variance. Then, we mapped each missense variant to the nearest gene and carried out pathway enrichment analysis. The Notch signaling pathway and Apoptosis pathway have been reported to be related to the formation of Alzheimer's disease.**

***Index Terms*—Image genetics, multiple regression tasks model, sparse learning model, Alzheimer's disease, whole genome association study.**

I. Introduction

ALZHEIMER disease (AD) is a chronic neurodegenerative disease and causes cognitive decline in elderly individuals. AD is pathologically defined by the presence of amyloid-β (Aβ) accumulation in amyloid plaques, tau aggregation in neurofibrillary tangles and brain atrophy caused by loss of neurons and synapses [1]. Apolipoprotein $E\varepsilon4$ allele (*APOE4*) is considered as a risk factor for AD [2]. *APOE* is highly expressed in astrocytes and disease-associated microglia (or DAMs) in the brain, implicating a role for these cell types in pathogenesis. Studies [3], [4] declared that the innate immune system and inflammation play a key role in formation of AD, so that those genes encoding triggering receptor expressed on myeloid cells 2 (*TREM2*), complement receptor 1 (*CR1*), *CD33* and inositol polyphosphate-5-phosphatase (*INPP5D*) are also potent risk factors. In addition to tumorigenesis, Notch signaling is also involved in determining the fates of neural and nonneural cells and alterations in proteolysis of the Notch by γ-secretase could play key role in the pathogenesis of AD [5]–[7]. Studies have indicated that apoptotic mechanisms could be instrumental in neurodegeneration in AD. Therefore, detailed and precise study of apoptotic mechanisms in AD is important for treating and preventing AD [8], [9].

Although lots of genetic mutations and biomolecules have been shown to be related to the AD, however, as a typical complex disease, the formation and progression mechanisms of AD are just seen in the tip of the iceberg. For example, each risk factor shows weak association signal with AD; more importantly, studies show that non-genetic risk factor (such as lifestyle, high blood pressure, obesity and so on) also plays an important role in disease pathogenesis [1]. Aiming at defining the progression of Alzheimer's disease, Alzheimer's Disease Neuroimaging Initiative (ADNI) researchers collect, utilize and share data, including functional and structural images of brain, genetics, blood biomarkers and so on [10]. With using the ADNI data, hundreds of studies have been carried out, including univariate analysis, multi-factor analysis and systems biology analysis, so that ADNI enriches human's knowledge about AD. Imaging genetics consider structural or functional neuroimaging data as endophenotype to pinpoint causal genetic variants, namely single nucleotide polymorphisms (SNP), related to AD [11]. Compared to traditional case-control studies, the functional and structural images derived from magnetic resonance imaging (MRI) and positron emission tomography (PET) can fill the gap between the genetic mutations and AD, so that imaging genetics have promising expectation in identification of pathogenic variants [12].

To consider the relationships between brain regions of interest (ROIs) and genomic variants, lots of early studies adopted

Manuscript received August 9, 2020; revised January 13, 2021; accepted June 23, 2021. Date of publication June 28, 2021; date of current version September 3, 2021. This work is supported in part by the National Natural Science Foundation of China under Grant 62062032, and in part by the Jiangxi Provincial Natural Science Fund under Grants 20192ACB21004 and 20204BCJL23035. *(Corresponding author: Xiong Li.)*

The authors are with the School of Software, East China Jiaotong University, Nanchang 330013, China (e-mail: lx_hncs@163.com; 984152245@qq.com; 956854931@qq.com; 1948573047@qq.com; 1091237067@qq.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2021.3093027, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2021.3093027

pairwise univariate correlation analysis methods which assume that input features are independent from each other and ROIs also are independent, resulting in ignoring the biological interactions between features. To address such issue, Yan *et al.* proposed a knowledge-guided sparse canonical correlation method (KG-SCCA) which borrows prior knowledge such as linkage disequilibrium and network structure. However, KG-SCCA also faces two important challenges: one is that the accuracy of the model depends heavily on the accuracy of prior information; the other is that the model is relatively complex and difficult to be applied to solve large-scale data sets (e.g. genome-wide analysis) [13]. In addition, multiple task regression model also has been widely used for studying imaging genetics [14], [15]. For the perspectives of feature selection and causal genetic variants identification, sparse techniques have been applied in multiple task regression model. In studies [16], [17], Lasso regression and Elastic Net have been applied to jointly evaluate the effects of variants on ROIs. However, Wang *et al.* believe that SNPs interact with each other through gene interaction or biological pathway, so that they proposed a G-SMuRFS (group-sparse multitask regression and feature selection) method based on group sparse model to describe group structural among SNPs [14]. Group sparse model not only provides rich biological meanings in image genetics, but also limits unnecessary combinations between SNPs. However, G-SMuRFS only furnishes a point estimate of the regression coefficients [15], Greenlaw *et al.* proposed a Bayesian method and implemented an R package *bgsmtr* for generating posterior means and credible intervals for each SNP. Although these group sparse model-based methods have certain advantages in some aspects, they still confront with three major challenges: firstly, note that different ROIs are influenced by different SNPs. Therefore, for a candidate SNPs association study, some causal SNPs may not be included in the candidate SNPs and more importantly, the multiple task regression model tends to select shared SNPs in different ROIs, so that it may neglect specific SNPs for some ROIs; Secondly, SNPs always have very weak association signal to phenotypes for complex diseases and the regression coefficients or credible intervals are difficult for researchers to determine which SNPs are actually related to AD. Consequently, there are hundreds of feature SNPs selected by G-SMuRFS and bgsmtr. However, solutions that contain too many SNPs are usually not credible. Lastly, due to difficulty in solving, the multiple task regression model is not available for high dimensional genetic data, let alone genome-wide studies.

In spite of the usefulness of L_0-based regularization, there is a steep computational price to pay when compared to search solutions satisfying optimality conditions. One strategy to alleviate this problem is by considering a larger family of estimators such as L_1 or L_2 norm regularization [18]. What is exciting is that Hazimeh *et al.* proposed a method based on coordinate descent and local combinatorial optimization for efficiently resolving L_0-regularized problems [19]. With using toolkit L0Learn [19], we design an L_0-based sparse model L0L2 for correlation analysis between SNPs and each ROI, which can address all these three challenges mentioned above. Then, we conduct a whole genome study with 808 samples and about 388 million SNPs for presenting the applicability of sparse model on large scale data.

II. Materials and Methods

A. Data Sources

This study will demonstrate the applicability of L_0-based sparse model for both image genetics and whole genome association study. Therefore, two kinds of datasets are involved:

Genetic and brain image measures datasets for image genetic study: One is simulated data (simulation dataset S1) with 632 subjects, 486 SNPs from 33 genes, 15 structural neuroimaging measures [15] and the other (simulation dataset S2) has 200 samples holding 50 SNPs from 10 groups and 50 structural neuroimaging measures [13]. These two datasets are mainly for correlation analysis between SNPs and ROIs.

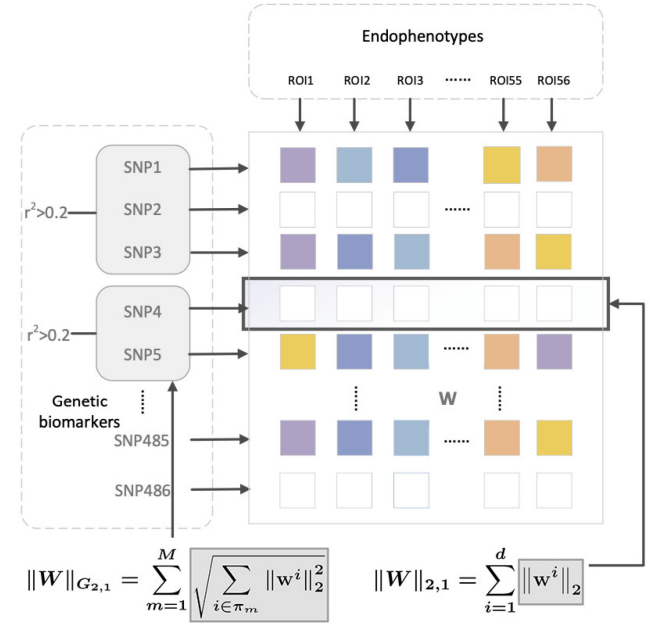
Genetic and phenotypes dataset for whole genome analysis: Data analyzed in this study were downloaded from the ADNI database (adni.loni.usc.edu). The ADNI was initially launched in 2003 and followed with three phases: ADNI-1, ADNI-2 and ADNI-GO. Whole-genome sequenced (WGS) at high coverage over 388 million SNPs was applied on ADNI1/GO/2 samples and genotyped using the Ilumina Omni 2.5 M BeadChip. In this study, there are 809 samples' variant call format (VCF) files recalled using the CASAVA pipelines. Out of 809 samples, 808 samples include 279 cognitive normal samples (CN), 483 mild cognitive impairment samples (MCI) and 46 Alzheimer's disease samples (AD) and the remaining one has no record about phenotypes, so that the unlabeled case has been discarded. Up-to-date information is presented on www.adni-info.org.

B. L_0-Based Sparse Model L0L2

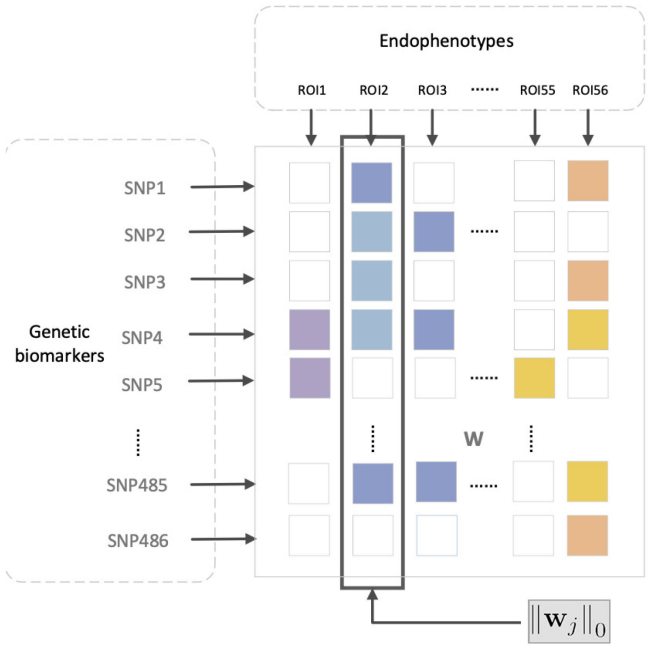
Given the SNP data of n subjects $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\} \subseteq \Re^d$ and c ROIs $\boldsymbol{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\} \subseteq \Re^c$ where d is the number of SNPs (feature dimensionality). There are three genotypes for each SNP: wild homozygote encoded as '0,' mutant heterozygote encoded as '1' and mutant homozygote encoded as '2'. All ROIs are continuous variable in different scales. To select feature SNPs from different groups while eliminating redundant SNPs in the same group, Wang *et al.* designed a sophisticated model shown in Formula (1), which includes $G_{2,1}$-norm and $L_{2,1}$-norm [14]. In the weight matrix $\mathbf{W}$, the entry $\mathbf{W}_{ij}$ measures the relative importance of the i-th SNP in predicting the response of the j-th ROI, and γ_1 and γ_2 are trade-off parameters of $G_{2,1}$-norm and $L_{2,1}$-norm. Fig. 1(a) approximately depicts the coefficients for determining which SNPs are important to which ROIs. For a multiple regression tasks model, $L_{2,1}$-norm tends to simultaneously shrink the single SNP's coefficients corresponding to different ROIs, which results in selecting shared features for all ROIs. Consequently, this characteristic of group sparse model may not suitable when different ROIs have different causal SNPs.

$$\min_{W} \sum_{i=1}^{n} \left\| \mathrm{W}^{\mathrm{T}}\mathrm{X} - \mathrm{Y} \right\|_F^2 + \gamma_1 \|\mathrm{W}\|_{G_{2,1}} + \gamma_2 \|\mathrm{W}\|_{2,1} \qquad (1)$$

where $\|W\|_{G_{2,1}} = \sum_{m=1}^{M} \sqrt{\sum_{i \in \pi_m} \|\mathbf{w}^i\|_2^2}$, $\|W\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{w}^i\|_2$ and w^i denotes the i-th row of $\mathbf{W}$, and $\{\pi_m\}_{m=1}^{M}$ denotes the group set in which the SNPs were divided into M groups according to gene or linkage disequilibrium.



(a)



(b)

Fig. 1. The shared feature SNPs selected by (a) group sparse model G-SMuRFS; (b) non-shared feature SNPs selected by L_0-based sparse model. The shades of color represent the degree of coefficient weights and the white grid indicates that the corresponding SNP has not been selected as a feature.

Here, to elevate the computational efficiency and regression accuracy of sparse model, we propose L_0-based sparse model for correlation analysis as shown in Formula (2).

$$\arg\min_{\mathbf{w_i} \in \mathbf{R}} \frac{1}{2} \|\mathbf{y_i} - \mathbf{X}\mathbf{w}_i\|_2^2 + \gamma_1 \|\mathbf{w}_i\|_0 + \gamma_2 \|\mathbf{w}_i\|_2^2 \quad (2)$$

where y_i denotes the i-th ROI, $\mathbf{w}_i$ represents the coefficients vector of all SNPs in predicting the response of the i-th ROI, γ_1 controls the number of feature SNPs with nonzeros coefficients and γ_2 controls the amount of shrinkage induced by L_2 regularization. The first component of Formula (2) is empirical risk of regression model, so as to ensure that the predicted value is as close as possible to the real value. The second component plays key role in identifying feature SNPs correlated to specific ROI, so that feature SNPs has nonzero values but other SNPs are all zeros. The third component is introduced to avoid overfitting.

In brief, the differences between Formula (2) and (1) are: First of all, our method is a single-task regression model which independently identifies specific feature SNPs for each ROI, resulting in feature SNPs more targeted; Secondly, although both L_0 and L_1 can achieve sparseness, L_0 can automatically exclude irrelevant SNPs by setting the coefficients of these SNPs to be zero without manual threshold setting. Because L_0-norm is introduced into the Formula (2), it becomes a non convex function. In [19], the non-convex problem of L_0 was clearly raised and a toolkit L0Learn was proposed to efficiently resolve this problem. The L0Learn adopts a coordinate cycle descent algorithm and local combination optimization algorithm to obtain optimal solutions and a spacer steps strategy is introduced to ensure the stability of the optimization process. The L0Learn toolkit is open-source in R/C++ and its speedups can reach up to three-fold when compared with state-of-the-art toolkit.

Note that we carried out correlation analysis between all SNPs and each ROI, respectively. Therefore, ROIs have different feature SNPs as shown in Fig. 1(b), which is significantly different from feature shared-based multiple tasks regression.

C. Evaluation Measures

We conduct two kinds of association studies: candidate SNPs-ROIs image genetics and whole genome variants-phenotypes association study. To declare the L_0-based method's advantages, root mean square error (RMSE) is applied for evaluating the regression error as defined in Formula (3). In the analysis of image genetic data, this measure RMSE is used to compare and analyze the prediction accuracy of the feature SNPs identified by the L0L2 and G-SMuRFS methods for all ROIs. The higher the prediction accuracy, the higher the correlation between these SNPs and ROIs.

$$RMSE = \sqrt{\frac{1}{\mathrm{n}} \sum_{i=1}^{n} (f_i - y_i)^2} \quad (3)$$

where f denotes the predicted value of the ROI and y denotes the real value of the ROI. where f denotes the predicted value of the ROI and y denotes the real value of the ROI.

In this study, we also use R^2 defined in Formula (4) to measure the proportions of ROIs or phenotype explained by feature SNPs. R^2 is used to measure the interpretability of SNPs for phenotypic variances in subsequent whole genome correlation analysis.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (4)$$

where $\overline{y}$ denotes the average of the ROI.

To demonstrate the applicability of L_0-based sparse model on high dimensional data, the issue of computational efficiency should be concerned. Therefore, we also compare the running time of method with group sparse L_0 method. Note that the configuration of the experimental computing platform are MacOS Catalina 10.15 operating system, 2.9 GHZ processor, Intel Quad-Core i7, and 16 GB random access memory.

D. K-Fold Cross-Validation

To verify if the obtained results are stable across training/test splits, *K*-fold cross-validation has been adopted to evaluate the performance of image genetics analysis methods more objectively. *K*-fold cross-validation reduces variance by averaging the training results of *K* different test sets, so the results are less sensitive to the data splits. Firstly, the original data were randomly divided into *K* splits without repeated sampling.Secondly, one split is selected as the test set and the remaining *K*-1 as the training set for model training. Thirdly, repeat the second step *K* times, so that each split has one chance as the test set and the rest as the training set. Finally, the average value of *K* results (e.g. RMSE) is calculated as the estimation of model accuracy. The setting of K value is usually based on the the size of the data set. When the scale of data is small, *K* can be set larger, so that the training set accounts for a larger proportion of the whole dataset, but at the same time, the number of training models also increases. When there is a large amount of data, *K* can be set smaller. In this study, we set *K* as 10.

III. Experimental Results

In order to illustrate the effectiveness and practicability of the proposed model L0L2, we carried out two kinds of experimental analysis in this section, namely, image genetics and whole genome correlation analysis. In the first part of this section, the advantages of the proposed L0L2 model in RMSE are demonstrated by comparative analysis of image genetics, which shows that the feature SNPs selected by our method can predict ROI more accurately. In the second part of this section, the whole genome univariate correlation analysis based on the *spearman* method is carried out. The third part of this section shows the whole genome multivariable correlation analysis based our L0L2 model. The purpose is to show that the multivariable model is more consistent with the pathogenesis of AD, and the identified susceptibility loci can explain phenotypic changes to a greater extent. Moreover, the real data analysis shows that L0L2 model can be effectively applied to higher dimensional data.

A. Image Genetic Simulation Study

For the first genetic and brain image measures simulation dataset S1 (632 subjects, 486 SNPs from 33 genes or groups, 15 ROIs), we compared L_0-based method L0L2 with G-SMuRFS on RMSE. To fairly compare the performance of regression accuracy, we applied 10-fold cross-validation strategy, which means that 9/10 datasets are used to train the model and test the model on the rest 1/10 and running this procedure 10 times.

Fig. 2 shows the RMSE results on six ROIs and other ROIs results shown Fig. S1 (**Supplementary**). These results show that L0L2 has better performance than G-SMuRFS, on average. It

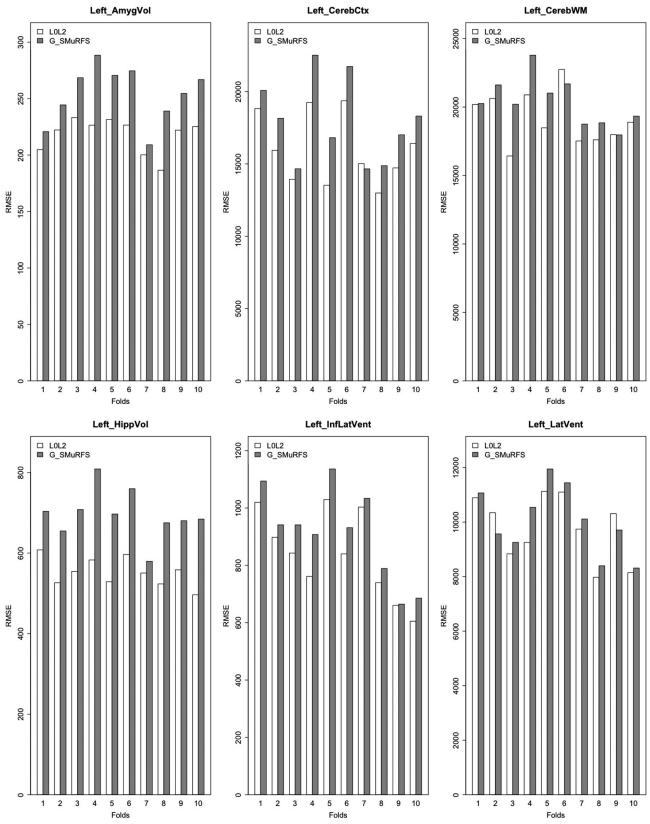


Fig. 2. 6 ROIs RMSE results of the simulation datasets S1.

means that considering image genetics as a multiple regression problem may be unsuitable, since different ROIs have different causal SNPs. More importantly, there is no main task among all these ROIs, since they are equal to each other and may not correlate with each other. In addition, all these ROIs have different scales, so that the learning rate may different.

Intuitively, each ROI should not be influenced by too many SNPs and some ROIs can not be affected by any causal SNPs in a candidate study. In Table I, each single ROI of the simulation dataset S1 corresponds to the feature SNPs identified by our L0L2 method. These results further validate that different ROIs have different causal SNPs. And, some ROIs have small set of feature SNPs, or even no feature SNPs (such as Left_HippVol). In addition, when selecting feature SNPs, our L_0-based method directly identifies non-zero coefficient SNPs as feature SNPs without manually setting the coefficient threshold or specifying the number of feature SNPs. This characteristic makes our computational method L0L2 more practical and makes geneticists more confident in our method. In contrast, G-SMuRFS selects too many SNPs with nonzero coefficients, which not only results in confusion for downstream analysis, but also includes noise in some degree.

Consequently, we believe that the correlations between SNPs and ROIs should be analyzed by single regression task model, but not multiple regression tasks model.

For an extensively comparison, we also carried out simulation study on the simulation dataset S2 (200 samples with 50 SNPs from 10 groups and 50 ROIs). Here, we also applied a 10-fold

TABLE I
THE FEATURE SNPS FOR EACH ROI OF THE SIMULATION DATASET S1

ROI name	Feature SNP S1	No. of SNPs
Left_AmygVol	rs4918282	1
Left_CerebCtx	rs4362, rs2986018, rs3811450, rs12378686, rs1473180, rs9426748, rs1433099, rs16871236, rs3798729, rs7748486, rs9393992, rs1150724, rs669556, rs10786972, rs10884402, rs1269918, rs661319, rs7095427, rs556349, rs689021	20
Left_CerebWM	rs2025935, rs1023024, rs11600875, rs4935774	4
Left_HippVol		0
Left_InfLatVent	rs685316	1
Left_LatVent	rs2025935, rs4877367, rs7025760, rs10422797, rs11601726, rs2018334, rs2182337, rs7938033, rs10787011, rs10884339, rs10884374, rs10884402, rs2243581, rs11006130	14
Left_EntCtx	rs642949, rs4631890	2
Left_Fusiform	rs12378686, rs11964334	2
Left_InfParietal	rs653765, rs8027998, rs11141889, rs12001404, rs12378686, rs3128519, rs3128521, rs3026883, rs6584307, rs1433099, rs1799898, rs17367504, rs6541003, rs11964334,rs7910584, rs11006133, rs2306604	17
Left_InfTemporal	rs12001404, rs3128521, rs12758257, rs212524, rs212525, rs213025, rs213028, rs6584307, rs6511720, rs11964334, rs11234495, rs475639, rs666682, rs677909, rs2756271, rs7073924, rs666004, rs689021, rs11006133, rs2306604	20
Left_MidTemporal	rs405509, rs11141889, rs12001404, rs3118846, rs3128521, rs12758257, rs212524, rs212525, rs9393992, rs11814111, rs6584766, rs7073924, rs3781827, rs666004, rs11006133, rs2306604, rs6503018	17
Left_Parahipp	rs10125534, rs10868609, rs1316489, rs2058882, rs4877367, rs212531, rs213023, rs213037, rs213039, rs471359, rs84853, rs17496723, rs4713379, rs9468690, rs10792820, rs713346, rs10884387, rs7903481, rs2276346	19
Left_PostCing	rs1473180, rs7036781, rs213025, rs16924159, rs11234495	5
Left_Postcentral	rs405509, rs10200967, rs729211, rs1014306, rs4878104, rs913782, rs213023, rs213052, rs471359, rs6584307, rs2239942, rs1433099, rs10501604, rs10501608,rs10884402, rs7897974, rs950809, rs676759, rs689021	19
Left_Precentral	rs17561	1

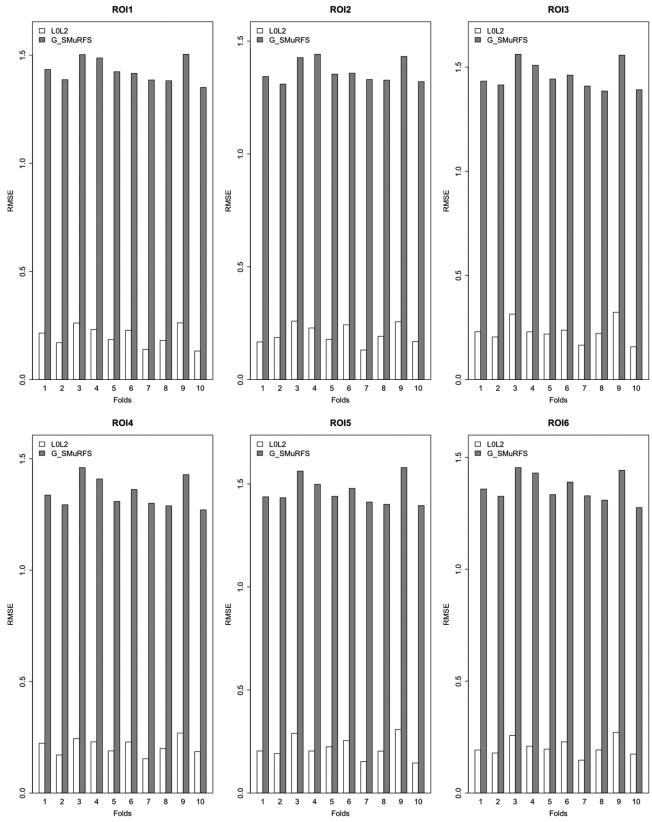


Fig. 3. 6 ROIs RMSE results of the simulation datasets S2.

cross-validation strategy. The first 6 ROIs regression results are shown in Fig. 3 and other results in Fig. S2 (Supplementary). From the results, we can find that our L0L2 method is significantly better than G-SMuRFS.

To test whether the 10-fold cross-validation results of L0L2 and G-SMuRFS are significantly different, we performed the following statistical analysis process:

1) For L0L2 and G-SMuRFS methods, RMSE means of each ROI in each fold were calculated separately;
2) Consider the RMSE means of L0L2 method and G-SMuRFS as paired samples and calculate the differences of paired samples;
3) Performing Shapiro-wilk test. If p-value$>$0.05, then the differences are following normal distribution, performing paired *t*-test, otherwise using paired samples Wilcoxon test;

The p-value of Shapiro-wilk test on simulation dataset S1 and simulation dataset S2 are 4.523e-06 and 0.008, which means that the differences do not follow normal distribution. Therefore, the paired samples Wilcoxon test is further applied and the p-value of them are 6.104e-05 and 7.79e-10, respectively. It means that for both datasets, the results of L0L2 and G-SMuRFS are significantly different. And, the median of the differences are all less than 0 ($-0.06\,185$ and $-1.17\,596$, respectively), indicating that the results of L0L2 method are significantly better than G-SMuRFS method.

TABLE II
THE RUNNING TIME IN SECONDS

	Simulation dataset S1	Simulation dataset S2
L0L2	105.521(s)	21.899(s)
G-SMuRFS	6177.190(s)	36.826(s)

To show the advantage of L_0-based method on computational efficiency, the running time of L0L2 and G-SMuRFS on these two simulation datasets are listed in Table II. From the results, we can find that as the scale of datasets grows, the advantage of L_0-based method on computational efficiency becomes more obvious. Thus, L_0-based method makes correlation study on high dimensional datasets possible.

B. Univariate Analysis for Whole Genome Analysis on ADNI

As a typical complex disease, it is also difficult to identify causal genetic variants or genes correlated to AD. Multiple loci locating in different genes interact with others through proteins interaction or pathways. It means that a single SNP may have very weak association signal to AD, so that the identified risk loci may only account for a small proportion of the heritability for AD. For examples, Ridge *et al.* concluded that only 33% of AD phenotypic variance can be described by common variants [20]; the APOEε4 allele only exists in about 50% of AD patients, suggesting that the interactions between other genetic variants or environmental factors may also contributing to risk for AD [21]. To uncover other unknown risk factors of AD, we conducted a whole genome analysis on ADNI.

We downloaded 809 samples from the ADNI. Of course, not all variants in whole genome will result in functional loss. The Ensembl Variant Effect Predictor (VEP) is a powerful genomic variants annotation and prioritization toolkit and it can determine the effect of SNPs on genes, transcripts, and protein sequence, as well as regulatory regions [22]. The VEP interprets variants based on reliable prior information, such as transcripts, clinical significance information, predictions of biophysical consequences of variants and so on. Inputting the coordinates of target variants and the amino acid substitution, we can find out genes and transcripts affected by target variants, consequence of target variants on the protein sequence, sorting intolerant from tolerant (SIFT) scores [23] for changes to protein sequence and so on. Note that SIFT algorithm predicts whether an nucleotide changes affects protein function with using prior knowledge, such as sequence homology and the physical properties of amino acids. Therefore, before carrying out downstream analysis, we used the VEP tool to annotate all 388 million SNPs. Then, we only kept these missense variants that pass the quality screening and sorting intolerant from tolerant SIFT score lower than 0.05. After that, more than 80 000 SNPs were left as candidate risk variants.

Out of 809 samples, there are 279 CNs, 483 MCIs, 46 ADs and the remaining one has no record about phenotypes, so that the unlabeled case has been discarded. In total, 808 samples carrying 80 556 SNPs in three categories were generated. In this study, we aim to identify risk factors which contribute the accumulation of risk for transition from CN to MCI or even AD. A patient with

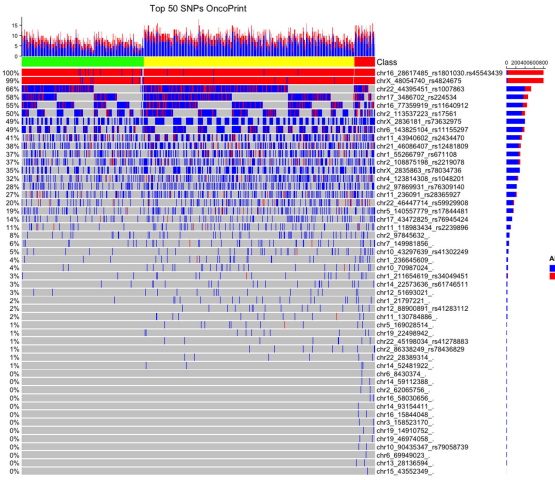


Fig. 4. The mutation frequencies of the top 50 SNPs. The column name is composed of chromosome number, physical position and SNP ID. When the SNP is unknown, it is denoted by '.'.

MCI shows a slight but noticeable and measurable decline in cognitive abilities, including memory and thinking skills [24]. MCI is at an increased risk of developing Alzheimer's or another dementia, so that in this study we rationally considered MCI as an intermediate state between CN and AD. Then, we labelled the status of CN as '0,' MCI as '1' and AD as '2'. Note that the label is ordered discrete variables. Therefore, the label can be considered as response variable and the L_0-based sparse model can be directly applied for the correlation analysis between SNPs and sample status (phenotypes).

First of all, we carried out a univariate association analysis with using *spearman* method. We find the rs224534 in chromosome 17 has the highest *spearman* value ($r = 0.12$, *p-value* $= 0.0006$) among these 80 556 SNPs. In order to uncover patterns hidden in genetic data across CN, MCI and AD samples, we used R package *ComplexHeatmap* [25] to visualize genomic alterations events. We sorted the SNPs according to the *spearman* value and depicted the mutation frequencies of the top 50 SNPs in Fig. 4. To uncover the link between SNPs and phenotype, each column in Fig. 4 represents a sample and each row represents the distribution of heterozygous mutant labeled as 'Heterozygote' and homozygous mutant 'Mutanthomozygote' across different kinds of samples. If there is a major pathogenic SNP that has a significant effect on the development of AD, we should be able to observe considerable mutation samples in phenotype MCI or AD and rarely in CN. However, in Fig. 4, these mutations are either nearly evenly distributed in different categories, or the mutation frequency is too low to have statistical effect, which indicates that there is no major SNP significantly correlated with phenotypes from the ADNI subjects.

C. Multivariate Analysis for Whole Genome Analysis on ADNI

Since every single SNP has minor effect on phenotypes, it is rational to consider multivariate correlation method, such as multivariate linear regression model. In this study, we proposed to apply our L0L2 method for correlation analysis between SNPs and phenotypes.

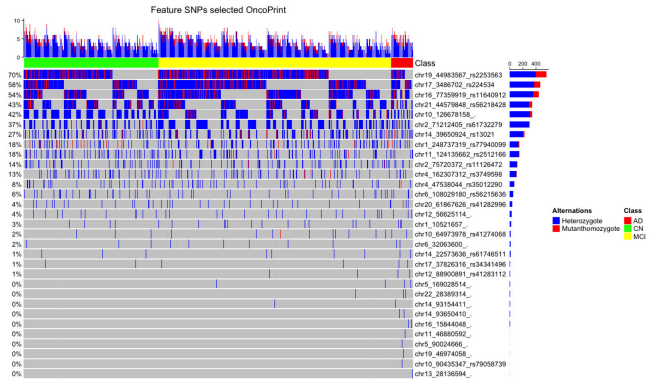


Fig. 5. The mutation frequencies of the 33 feature SNPs selected by L0L2 method. The column name is composed of chromosome number, physical position and SNP ID. When the SNP is unknown, it is denoted by '.'.

TABLE III
4 PATHWAYS GENERATED BY ENRICHMENT ANALYSIS

Name	P-value	Odds Ratio	Involved gene
Notch signaling pathway	0.002840	25.25	*NCOR2*, *CTBP2*
Tyrosine metabolism	0.05777	16.84	*PNMT*
Fat digestion and absorption	0.06553	14.78	*LIPF*
Apoptosis	0.02319	8.48	*DFFA*, *BIRC7*

With using L0L2 method, there are 33 feature SNPs with nonzero coefficients, and their mutation frequencies are depicted in Fig. 5. Similarly, none of them have main effect on phenotypes. However, note that the R^2 of these 33 feature SNPs reached up to 40.1%. It means that more than 40% of total AD phenotypic variance can be explained by the collective effect of these feature SNPs.

Next, we will examine which genes are affected by these feature SNPs and which biological pathways these genes participate in. First of all,, we mapped these missense variants to nearest gene, so that *ADAMTS18, CTBP2, TRPV1, AP001631.10, OR8G5, ZNF180, PNN, OR2T34, EVA1A, DFFA, ANKRD53, ATP10D, FSTL5, SLC39A5, SCML4, JMJD1C, TNXB, KITLG, SPDL1, TTC28, RIN3, BIRC7, MOAP1, TRAV24, MYH11, LRP4, PNMT, GPR98, PNMAL1, LIPF, LNX2, TOM1* and *NCOR2* are located. After that, we used these genes for enrichment analysis. Enrichment analysis computes enrichment based on gene-set libraries which are derived from prior knowledge and ranks functional terms from gene-set libraries by comparing the statistical results(e.g. Fisher exact test). Then, each set of genes associated with a functional term such as biological pathway is generated. Note that Enrichr [26] provides a friendly interface of KEGG (Kyoto Encyclopedia of Genes and Genomes) 2019 HUMAN. Consequently, 4 pathways have been derived and the detailed results are listed in Table III.

NCOR2 and *CTBP2* are involved in the Notch signaling pathway which is involved in determining the fates of neural and nonneural cells and alterations in proteolysis of the Notch by γ-secretase could play key role in the pathogenesis of AD [5]–[7]. *DFFA* and *BIRC7* are involved in the Apoptosis pathway which is important for treating and preventing AD [8][9]. Although Tyrosine metabolism pathway and Fat digestion and absorption pathway have no evidence for them related to AD in previous studies, pilot experiments could be designed for understanding their functions in AD.

IV. DISCUSSION AND CONCLUSION

In this study, through the results (the number of feature SNPs, the regression accuracy and running time) of SNPs and ROIs correlation analysis, we concluded that multiple regression tasks model may not be suitable for image genetics studies. The key reasons are as follows:

1) There is no main regression task among these ROIs. Although there may be interactions between ROIs, we can not borrow any knowledge from specific ROI to other ROIs. Therefore, it may be different from multiple regression tasks problem.
2) Each ROI may be influenced by different SNPs. In addition, the causal SNPs may not exist in the candidate SNPs. It means that the shared features selected by multiple regression tasks model just are just the fitted solutions without reasonably biological meanings.
3) Although there are lots of available normalization techniques, the significant differences in the scale of ROIs may also influence the findings of correlation analysis. Besides, the scale of different ROIs may vary greatly, so that the optimal learning rates between ROIs may be different.

Based on above reasons, we separately handle each single ROI to all SNPs by using L_0-based method. Through the extensively experimental results of image genetics and whole genome analysis on ADNI, we find that there are three main advantages of L_0-based method. Firstly, the zero coefficients of unrelated SNPs can be excluded without manual threshold setting. At least, it will help biology researchers to make decisions with more confidence. Secondly, the smaller number of feature SNPs selected by L_0-based method has better regression accuracy than other methods, which means that our method can effectively exclude redundant SNPs and select the most representative feature SNPs at the same time. Lastly, L_0-based model can be efficiently resolved by the L0L2 toolkit and as the scale of datasets grows, the computational efficiency gets more obvious.

From the results of univariate analysis on whole genome association study on ADNI, we find that as a typical complex disease, a single SNP has minor effect or even less on phenotype variants. With the advantages of L_0-based method, we also carried out a multivariate analysis for considering interactions between SNPs. The real experiment results show that the feature SNPs selected by our method not only explains more than 40% of total phenotype variance, but also enriches in two important pathways reported by previous studies. These findings are mainly due to: (1) fully application of the VEP tool to annotate all SNPs and exclude lots of non-functional mutations; (2) simple but effect L_0-based method can depict the interactions landscape in high dimensional genetic datasets.

Although our method shows potent performance and derives some interesting findings, several issues should to be resolved

in future work. For example, we only kept the missense variants, which ignores mutations located in the upstream and downstream regulators; Other unknown risk factors such as copy number variants and methylation also should be addressed to enhance the explanation of phenotype variances.

DATA AND CODE AVAILABILITY

Code is publicly available at: https://github.com/XiongLi2016/L0L2_image_genetics. And, the real brain image and genetic data can be accessed in http://adni.loni.usc.edu/data-samples/access-data/.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGMENT

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. 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, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; 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 ([1]). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

[1] [Online]. Avilable: www.fnih.org

REFERENCES

[1] C. M. Henstridge, B. T. Hyman, and T. L. Spires-Jones, "Beyond the neuron-cellular interactions early in Alzheimer disease pathogenesis," *Nature Rev. Neurosci.*, vol. 20, no. 2, pp. 94–108, 2019.
[2] E. H. Corder *et al.*, "Gene dose of apolipoprotein e type 4 allele and the risk of Alzheimer's disease in late onset families," *Sci.*, vol. 261, no. 5123, pp. 921–923, 1993.
[3] L. J. Van Eldik *et al.*, "The roles of inflammation and immune mechanisms in Alzheimer's disease," *Alzheimer's Dementia: Transl. Res. Clin. Interv.*, vol. 2, no. 2, pp. 99–109, 2016.
[4] K. D. van der Willik *et al.*, "Balance between innate versus adaptive immune system and the risk of dementia: A population-based cohort study," *J. Neuroinflamm.*, vol. 16, no. 1, pp. 1–9, 2019.
[5] H.-N. Woo, J.-S. Park, A.-R. Gwon, T. V. Arumugam, and D.-G. Jo, "Alzheimer's disease and notch signaling," *Biochem. Biophysical Res. Commun.*, vol. 390, no. 4, pp. 1093–1097, 2009.
[6] J. L. Lasky and H. Wu, "Notch signaling, brain development, and human disease," *Pediatr. Res.*, vol. 57, no. 7, pp. 104–109, 2005.
[7] R. Kopan and A. Goate, "A common enzyme connects notch signaling and alzheimer's disease," *Genes Develop.*, vol. 14, no. 22, pp. 2799–2806, 2000.
[8] D. W. Dickson, "Apoptotic mechanisms in alzheimer neurofibrillary degeneration: Cause or effect," *J. Clin. Investig.*, vol. 114, no. 1, pp. 23–27, 2004.
[9] C. Behl, "Apoptosis and Alzheimer's disease," *J. Neural Transmiss.*, vol. 107, no. 11, pp. 1325–1344, 2000.
[10] M. W. Weiner *et al.*, "The Alzheimer's disease neuroimaging initiative: Progress report and future plans," *Alzheimer's Dement.*, vol. 6, no. 3, pp. 202–211e7, 2010.
[11] L. Shen *et al.*, "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the adni cohort," *Neuroimage*, vol. 53, no. 3, pp. 1051–1063, 2010.
[12] Z. Xu *et al.*, "Imaging-wide association study: Integrating imaging endophenotypes in GWAS," *Neuroimage*, vol. 159, pp. 159–169, 2017.
[13] J. Yan *et al.*, "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm," *Bioinformatics*, vol. 30, no. 17, pp. i564–i571, 2014.
[14] H. Wang *et al.*, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the adni cohort," *Bioinformatics*, vol. 28, no. 2, pp. 229–237, 2012.
[15] K. Greenlaw, E. Szefer, J. Graham, M. Lesperance, F. S. Nathoo, and A. D. N. Initiative, "A Bayesian group sparse multi-task regression model for imaging genetics," *Bioinformatics*, vol. 33, no. 16, pp. 2513–2522, 2017.
[16] O. Kohannim *et al.*, "Discovery and replication of gene influences on brain structure using lasso regression," *Front. Neurosci.*, vol. 6, 2012, Art. no. 115.
[17] O. Kohannim *et al.*, "Predicting temporal lobe volume on mri from genotypes using l 1-l 2 regularized regression," in *Proc. 9th IEEE Int. Symp. Biomed. Imag.*, 2012, pp. 1160–1163.
[18] A. Patrascu, and I. Necoara, "Random coordinate descent methods for l0 regularized convex optimization," *IEEE Trans. Automatic Control*, vol. 60, no. 7, pp. 1811–1824, 2015.
[19] H. Hazimeh and R. Mazumder, "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms," *Operations Research*, vol. 68, no. 5, pp. 1517–1537, 2020.
[20] P. G. Ridge, S. Mukherjee, P. K. Crane, and J. S. Kauwe, "Alzheimer's disease: Analyzing the missing heritability," *PLoS One*, vol. 8, no. 11, 2013, Art no. e79771.
[21] C. M. Karch, C. Cruchaga, and A. M. Goate, "Alzheimer's disease genetics: From the bench to the clinic," *Neuron*, vol. 83, no. 1, pp. 11–26, 2014.
[22] W. McLaren *et al.*, "The ensembl variant effect predictor," *Genome Biol.*, vol. 17, no. 1, pp. 1–14, 2016.
[23] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003.
[24] D. Müller-Gerards *et al.*, "Subjective cognitive decline, apoe 4, and incident mild cognitive impairment in men and women," *Alzheimer's Dementia: Diagnosis, Assessment Dis. Monit.*, vol. 11, pp. 221–230, 2019.
[25] Z. Gu, R. Eils, and M. Schlesner, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data," *Bioinformatics*, vol. 32, no. 18, pp. 2847–2849, 2016.
[26] M. V. Kuleshov *et al.*, "Enrichr: A comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W90–W97, 2016.