BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 2 2012, pages 229–237
doi:10.1093/bioinformatics/btr649

Genetics and population analysis

Advance Access publication December 6, 2011

Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort

Hua Wang[1,†], Feiping Nie[1,†], Heng Huang[1,*], Sungeun Kim[2], Kwangsik Nho[2], Shannon L. Risacher[2], Andrew J. Saykin[2], Li Shen[2,*]; For the Alzheimer's Disease Neuroimaging Initiative[‡]

[1]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
[2]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, 950 W. Walnut St, R2 E124F, Indianapolis, IN 46202, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Recent advances in high-throughput genotyping and brain imaging techniques enable new approaches to study the influence of genetic variation on brain structures and functions. Traditional association studies typically employ independent and pairwise univariate analysis, which treats single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) as isolated units and ignores important underlying interacting relationships between the units. New methods are proposed here to overcome this limitation.

Results: Taking into account the interlinked structure within and between SNPs and imaging QTs, we propose a novel Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) method to identify quantitative trait loci for multiple disease-relevant QTs and apply it to a study in mild cognitive impairment and Alzheimer's disease. Built upon regression analysis, our model uses a new form of regularization, *group* $\ell_{2,1}$*-norm* ($G_{2,1}$*-norm*), to incorporate the biological group structures among SNPs induced from their genetic arrangement. The new $G_{2,1}$*-norm* considers the regression coefficients of all the SNPs in each group with respect to all the QTs together and enforces sparsity at the group level. In addition, an $\ell_{2,1}$-norm regularization is utilized to couple feature selection across multiple tasks to make use of the shared underlying mechanism among different brain regions. The effectiveness of the proposed method is demonstrated by both clearly improved prediction performance in empirical evaluations and a compact set of selected SNP predictors relevant to the imaging QTs.

Availability: Software is publicly available at: http://ranger.uta.edu/%7eheng/ imaging-genetics/

Contact: heng@uta.edu; shenli@iupui.edu

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 12, 2011; revised on November 1, 2011; accepted on November 17, 2011

1 INTRODUCTION

Imaging genetics is an emergent transdisciplinary research field, where the associations between genetic variations and imaging measures as quantitative traits (QTs) or continuous phenotypes are evaluated. Compared to case–control status, the QTs have increased statistical power and are closer to the underlying biological etiology of the disease making it easier to identify underlying genes (Braskie *et al.*, 2011; Potkin *et al.*, 2009; Shen *et al.*, 2010; Stein *et al.*, 2010; Yip and Lange, 2011; Zhan *et al.*, 2011). Genome-wide association studies (GWAS) have been increasingly performed to correlate high-throughput single nucleotide polymorphism (SNP) data to large-scale image data. While many studies employed a hypothesis-driven approach by making significant reduction in one or both data types (Glahn *et al.*, 2007), some recent studies examined these associations at the whole genome entire brain level (Shen *et al.*, 2010; Stein *et al.*, 2010). Pairwise univariate analysis was typically used in traditional association studies to quickly provide important association information between SNPs and QTs. However, it treated the SNPs and the QTs as independent and isolated units, and therefore the underlying interacting relationships between the units might be lost. Multivariate methods to examine joint effect of multi-locus genotype on a single phenotype were studied in general genetic association studies (Ballard *et al.*, 2010; Wu *et al.*, 2010) as well as several recent imaging genetic studies (Bralten *et al.*, 2011; Hibar *et al.*, 2011). This paradigm did not consider the relationship between interlinked brain phenotypes and thus still had limited power in revealing complex imaging genetic associations. In this work, taking into account the interrelated structure within and between SNPs and QTs, we propose a new framework for effectively identifying quantitative trait loci, which addresses the following challenges in imaging genetics association study.

First, traditional association studies consider all the SNPs evenly distributed and assess each SNP individually. However, certain SNPs are naturally connected via different pathways. Multiple SNPs from one gene often jointly carry out genetic functionalities. Moreover,

linkage disequilibrium (LD) (Barrett *et al.*, 2005) describes the non-random association between alleles at different loci, through which the SNPs in high LD are linked together in meiosis. Thus, instead of treating SNPs in an isolated manner, it would be beneficial to exploit the group structure among SNPs.

Second, because the functionality of the human brain typically involves more than one cerebral component, investigating each individual regional brain phenotype separately will inevitably lose the interacting relationships between them. For example, the brain's episodic memory network, including medial temporal lobe (MTL) structures, medial and lateral parietal, and prefrontal cortical areas, are normally engaged together during episodic recall (Walhovd *et al.*, 2010). In addition, accurate prediction of disease status and progression are typically implicated by multiple brain regions coupled with other biomarkers (Hinrichs *et al.*, 2011; Zhang *et al.*, 2011). Therefore, jointly analyzing all the imaging phenotypes via one single integral regression model is desirable to elucidate the shared mechanism that may be hidden otherwise.

By recognizing the interrelated nature of these genotypes and phenotypes, in this study, we propose a novel Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) method to identify quantitative trait loci in a mild cognitive impairment (MCI) and Alzheimer's disease (AD) study using a few important imaging QTs relevant to AD. We consider each SNP as a *feature* and each QT as a *response variable* (i.e. a *learning task*), and formulate a multitask regression framework including multiple features (SNPs) and multiple responses (QTs). Our goal is to reveal the relationships between these genetic features and imaging phenotypes.

The proposed model consists of three major components. First, it is built upon regression analysis due to the continuous responses of the imaging phenotypes. As a result, the regression coefficients assess the relationships between SNPs and QTs. Second, in order to address the group-wise association among SNPs, inspired by group Lasso (Yuan and Lin, 2006), we propose a new form of regularization, called as *group $\ell_{2,1}$-norm ($G_{2,1}$-norm) regularization*, in which the coefficients of the SNPs within a pre-defined group, with respect to all the QTs, are penalized as a whole via ℓ_2-norm, while ℓ_1-norm is used to sum up the group-wise penalties to enforce sparsity between groups (Tibshirani, 1996). The latter is important because in reality only a small fraction of genotypes are related to a specific phenotype. Moreover, with sparsity, outliers and irrelevant associations are inherently removed. Lastly, through enforcing $\ell_{2,1}$-norm regularization, feature selection becomes an integrated procedure across multiple learning tasks (Argyriou *et al.*, 2007; Obozinski *et al.*, 2006), such that the interrelationships among different imaging phenotypes are leveraged. Note that the proposed $G_{2,1}$-norm and the enforced $\ell_{2,1}$-norm couple a set of learning tasks together such that the regression analysis can be carried out jointly across all the QTs, whereas Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006) perform regression analysis separately, one task at a time.

We apply the proposed G-SMuRFS method to the Alzheimer's disease neuroimaging initiative (ADNI) cohort (Weiner *et al.*, 2010) for identifying quantitative trait loci (QTLs) in MCI and AD using a set of imaging phenotypes known to be relevant to AD. Our empirical results yield not only clearly improved prediction performance in all test cases, but also a compact set of SNP predictors relevant to the imaging genotypes that are in accordance with prior studies.

2 MATERIALS AND DATA SOURCES

Both SNP and structural magnetic resonance imaging (MRI) data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). One goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. Following a prior study (Shen *et al.*, 2010), 733 non-Hispanic Caucasian participants were included in this study.

2.1 SNP genotyping and group information

The SNP data (Saykin *et al.*, 2010), used in this study, were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). Among all SNPs, only SNPs, belonging to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of June 10, 2010, were selected after the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy–Weinberg equilibrium test, (v) marker removal by the minor allele frequency and (vi) population stratification. As the second pre-processing step, the quality-controlled SNPs were imputed using the MaCH software (Li *et al.*, 2010) to estimate the missing genotypes. After that, the Illumina annotation information based on the Genome build 36.2 was used to select a subset of SNPs, belonging to the top 40 AD candidate genes (Bertram *et al.*, 2007).

The above procedure yielded 1224 SNPs from 37 genes. For the remaining three genes, no SNPs were available on the genotyping chip. The genes and the number of their SNPs are shown in Figure 1. The ranking of the AlzGene database is based on SNPs instead of genes. As a result, most of the SNPs from these genes (Fig. 1) might be irrelevant to AD, while a small fraction of them could be risk factors for the disease and be associated with our intermediate imaging traits. Our task is to identify the SNPs in these 37 genes that predict important imaging QTs.

A straightforward observation from Figure 1 shows that the SNPs are naturally divided into groups upon their belonging genes. This grouping structure of SNPs, though conveying important biological information, is seldom utilized in previous association studies that consider every SNP equally and investigate their genetic effects on imaging phenotypes separately. In this work, as one of the contributions, we aim to make use of the grouping information of

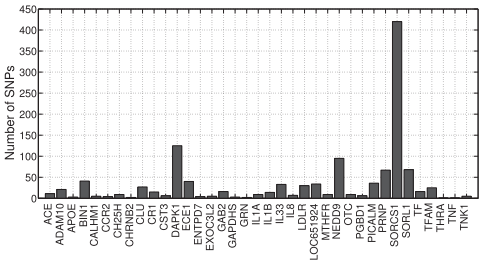

Fig. 1. Top 37 AD risk factor genes used in this study and the numbers of their SNPs.

SNPs in our learning model so as to achieve more lucid relationships between SNPs and neuroimaging phenotypes.

Besides grouping SNPs by genes, an alternative method could be based on LD. Through estimating non-random association of alleles at different loci (e.g. using pairwise correlation coefficients r^2, as shown in Figure 2), the relationships between SNPs in terms of genetic linkage are established. For example, the group structure can be clearly observed in Figure 2, where a group is defined as a block of SNPs whose pairwise $r^2 \geq 0.2$. As a result, we have 185 groups comprising 1029 SNPs, with each of the remaining 195 SNPs being isolated by itself.

In this study, we consider grouping SNPs by both genes and LD correlation coefficients r^2.

2.2 MRI analysis and extraction of imaging genotypes

Two widely employed automated MRI analysis techniques were used to process and extract imaging genotypes across the brain from all baseline scans of ADNI participants as previously described (Shen *et al.*, 2010). First, voxel-based morphometry (VBM) (Ashburner and Friston, 2000) was performed to define global gray matter (GM) density maps and extract local GM density values for target regions. Second, automated parcellation via FreeSurfer V4 (Fischl *et al.*, 2002) was conducted to define volumetric and cortical thickness values for regions of interest (ROIs) and to extract total intracranial volume (ICV). Further information is available in (Shen *et al.*, 2010). While a complete investigation of all VBM and FreeSurfer measures is an interesting future direction, this study is focused on a subset of these measures to test the proposed methods. Ten VBM (GM density) measures and 12 FreeSurfer measures (thickness/volume), which are known to be related to AD, are selected as QTs for identifying QTLs. These QTs are extracted from roughly matching ROIs with VBM and FreeSurfer. Table 1 shows the description of these QTs and Figure 3 maps some of these ROIs in the brain space. All these measures were adjusted for the baseline age, gender, education, handedness and baseline ICV using the regression weights derived from the healthy control participants.

3 METHODS

In this section, we first systematically develop our computational models to explore the associations between SNPs and imaging phenotypes. As illustrated in Figure 4, our method mainly addresses the group structure of genetic markers and joint learning across all the imaging endophenotypes,

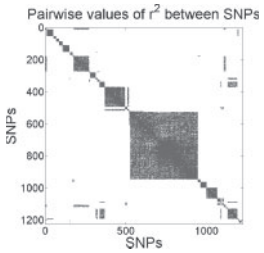

Fig. 2. Pairwise LD correlation coefficients ($r^2 > 0.2$ in blue) among the 1224 SNPs used in this study. The SNPs clearly form groups.

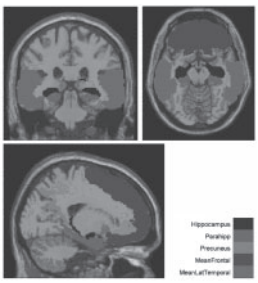

Fig. 3. VBM ROIs used in this study are mapped onto a brain.

Table 1. QTs from 'matching' ROIs: the volumetric/thickness measures (FreeSurfer) and GM density measures (VBM)

Volume/Thickness (ID and ROI)		GM Density (ID and ROI)	
LHippVol RHippVol	Volume of hippocampus	LHippocampus Rhippocampus	Hippocampus
LEntCtx LParahipp REntCtx RParahipp	Thickness of entorhinal cortex and thickness of parahippocampal gyrus	LParahipp RParahipp	Parahippocampal gyrus
LPrecuneus RPrecuneus	Thickness of precuneus	LPrecuneus RPrecuneus	Precuneus
LMeanFront RMeanFront	Mean thickness of caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole	LMeanFrontal RMeanFrontal	Inferior frontal operculum, inferior orbital frontal gyrus, inferior frontal triangularis, medial orbital frontal gyrus, middle frontal gyrus, middle orbital frontal gyrus, superior frontal gyrus, medial superior frontal gyrus, superior orbital frontal gyrus, rectus gyrus, rolandic operculum and supplementary motor area
LMeanLatTemp RMeanLatTemp	Mean thickness of inferior temporal, middle temporal, and superior temporal gyri	LMeanLatTemporal RMeanLatTemporal	Inferior temporal gyrus, middle temporal gyrus and superior temporal gyrus

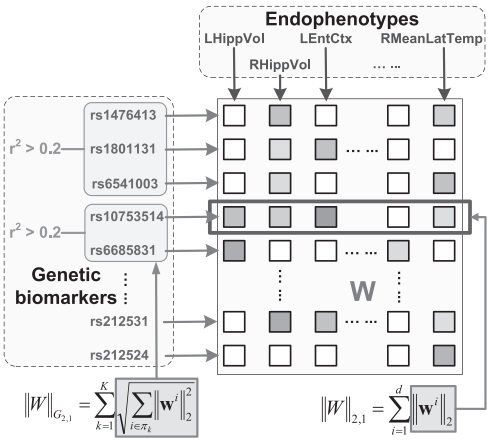

Fig. 4. Illustration of the proposed G-SMuRFS method. We incorporate the group structural information of the genetic markers through a new group $\ell_{2,1}$-norm regularization ($||\mathbf{W}||_{G_{2,1}}$), and enforce $\ell_{2,1}$-norm regularization ($||\mathbf{W}||_{2,1}$) to jointly select prominent SNPs across all endophenotypes.

such that the learned regression model has better prediction performance and the selected SNPs are more biologically meaningful. After that, we provide a new efficient algorithm to solve the proposed new multitask regression and feature selection objective, followed by the rigorous algorithm analysis to prove its correctness and convergence.

Throughout this article, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. Given a matrix $\mathbf{M}=(m_{ij})$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. The Frobenius norm and $\ell_{2,1}$-norm (also called as $\ell_{1,2}$-norm) of a matrix are defined as $||\mathbf{M}||_F=\sqrt{\sum_i||\mathbf{m}^i||_2^2}$ and $||\mathbf{M}||_{2,1}=\sum_i\sqrt{\sum_j m_{ij}^2}=\sum_i||\mathbf{m}^i||_2$, respectively.

3.1 G-SMuRFS

To explore the associations between SNPs and continuous imaging phenotypes, the linear (least square) regression (LR) is a standard approach. To avoid overfitting and increase numerical stability, the ridge regression (RR) is a better option. Given the SNP data of the ADNI participants $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}\subseteq\Re^d$ and the selected imaging phenotypes $\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}\subseteq\Re^c$, where n is the number of participants (sample size), d is the number of SNPs (feature dimensionality) and c is the number of imaging phenotypes (tasks), the RR is designed to solve:

$$\min_{\mathbf{W}}\sum_{i=1}^{n}||\mathbf{W}^T\mathbf{x}_i-\mathbf{y}_i||^2+\gamma\sum_{i=1}^{d}||\mathbf{w}^i||^2, \tag{1}$$

where the entry w_{ij} of the weight matrix $\mathbf{W}$ measures the relative importance of the i-th SNP in predicting the response of the j-th imaging phenotype, and $\gamma>0$ is a trade-off parameter.

However, the RR model in Equation (1) suffers from a number of problems when applied to evaluation of the imaging genetic associations. First, the weight matrix $\mathbf{W}$ is not sparse, therefore all the SNPs are involved in the prediction of imaging phenotype responses. However, among numerous SNPs, only a small fraction of them are relevant to specific imaging QTs. Thus, it is desirable to select only relevant SNPs for more accurate prediction. Second, similar to LR, the tasks in the RR regression model are decoupled and each of them can be learned separately. As a result, the information of underlying interacting relationships between the brain regions are ignored, which, though, are essential to brain functionalities. Finally, the rows of $\mathbf{W}$ are equally treated in the RR model, which implies that the underlying structures among these SNPs are overlooked. However, it is generally believed that many SNPs are genetically linked. In order to tackle these difficulties, we propose a novel G-SMuRFS method to exploit the interrelated structures within and between the genotypes and phenotypes.

3.1.1 Group-sparsity for genetic association The objective of RR model in Equation (1) uses Frobenious norm for regularization, which penalizes all the coefficients in a flat manner thereby all the SNPs are evenly treated. However, SNPs on the same chromosome with close distance tend to be inherited together and correlated with each other. For example, as shown in Figure 4, the pairwise LD correlation coefficients r^2 between 'rs1476413', 'rs1801131' and 'rs6541003' are >0.2, thus they are more homogeneous and should be considered together when we predict the responses of the imaging QTs. Motivated by sparse learning, such as Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006), we propose a new form of regularization as follows:

$$\min_{\mathbf{W}}\sum_{i=1}^{n}||\mathbf{W}^T\mathbf{x}_i-\mathbf{y}_i||_2^2+\gamma\sum_{k=1}^{K}\sqrt{\sum_{i\in\pi_k}\sum_{j=1}^{c}w_{ij}^2}, \tag{2}$$

where the SNPs, i.e. features, are partitioned into K groups $\Pi=\{\pi_k\}_{k=1}^K$, such that $\{\mathbf{w}^i\}_{i=1}^{m_k}\in\pi_k$ are genetically linked, and m_k is the number of SNPs in π_k. Two types of genetic links are used here to group SNPs: (i) SNPs are naturally divided into groups based on their belonging or nearest genes. (ii) SNPs are grouped by thresholding the pairwise LD correlation coefficients r^2, e.g. in this work, the neighboring SNPs whose $r^2\geq 0.2$ form a group.

Without loss of generality, $\{\pi_k\}_{k=1}^K$ are ordered and concatenated. Denote $\mathbf{W}=\begin{bmatrix}\mathbf{W}^1\\ \cdots\\ \mathbf{W}^K\end{bmatrix}$, where $\mathbf{W}^k\in\mathbb{R}^{m_k\times c}(1\leq k\leq K)$, we can write Equation (2) as following:

$$\min_{\mathbf{W}}\sum_{i=1}^{n}||\mathbf{W}^T\mathbf{x}_i-\mathbf{y}_i||_2^2+\gamma\sum_{k=1}^{K}||\mathbf{W}^k||_F, \tag{3}$$

which can be written in matrix form as following:

$$\min_{\mathbf{W}}||\mathbf{W}^T\mathbf{X}-\mathbf{Y}||_F^2+\gamma||\mathbf{W}||_{G_{2,1}}, \tag{4}$$

where $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]$, $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]$, and $||\cdot||_{G_{2,1}}$ is our proposed *group $\ell_{2,1}$-norm ($G_{2,1}$-norm)* of a matrix with respect to a partition Π and defined as:

$$||\mathbf{W}||_{G_{2,1}}=\sum_{k=1}^{K}\sqrt{\sum_{i\in\pi_k}\sum_{j=1}^{c}w_{ij}^2}=\sum_{k=1}^{K}||\mathbf{W}^k||_F\ . \tag{5}$$

Note that the $G_{2,1}$-norm defined above is different from the regularization term in group Lasso. Given a partition of the features, the group Lasso enforces group-wise sparsity for each learning task separately, whereas the $G_{2,1}$-norm defined in Equation (5) penalizes the regression coefficients of a group of features across all the learning tasks jointly. As a result, the biological group-level structural information among SNPs are incorporated into our multi-task learning model.

Moreover, because the ℓ_1-norm across all the group-wise penalties are used in $G_{2,1}$-norm, similar to Lasso and group Lasso, sparsity is enforced among biological groups. This is important in identifying relevant genotypes for specific phenotypes, because only a small fraction of SNPs are related to certain imaging phenotypes. From the perspective of sparsity learning, the Lasso and group Lasso have flat sparsity, the $\ell_{2,1}$-norm has structured sparsity, and the $G_{2,1}$-norm has structured sparsity across feature groups.

3.1.2 Individual structured sparsity for joint feature selection Although the objective in Equation (4) takes into account the group structure of the SNP data through the proposed $G_{2,1}$-norm, the feature selection across tasks are still not completely addressed, because $G_{2,1}$-norm penalizes the coefficients flatly within each group of SNPs. To be more specific, within a given group, say π_k, Frobenious norm $||\mathbf{W}^k||_F$ is used, which is the same as ridge regression that uses Frobenious norm over the whole projection matrix $\mathbf{W}$. In an important group, certain features could be irrelevant; on the other

hand, in a less important group, some features could be significant to tasks. Thus, we enforce additional structured sparsity to our learning model for jointly selecting features across multiple tasks via a $\ell_{2,1}$-norm regularization (Argyriou *et al.*, 2007; Lee *et al.*, 2010; Obozinski *et al.*, 2006; Puniyani *et al.*, 2010):

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}^i||_2, \tag{6}$$

which can be concisely rewritten in matrix form as:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 ||\mathbf{W}||_{G_{2,1}} + \gamma_2 ||\mathbf{W}||_{2,1} \ . \tag{7}$$

In Equation (7), the first term measures the regression loss. The second term couples all the regression coefficients of a group of features over all the c tasks together, which incorporates the grouping information on features (SNPs) due to the genetic linkage. Finally, the third term penalizes all c regression coefficient of each individual feature as whole to select features across multiple learning tasks.

We call Equation (7) as G-SMuRFS method with illustration in Figure 4.

3.2 A new efficient optimization algorithm

Because the number of genetic markers can be very large, we need an efficient algorithm to solve Equation (7). Existing algorithms usually reformulate such sparsity problem as a second-order cone programming (SOCP) or semidefinite programming (SDP) problem, which can be solved by interior point method or the bundle method. However, solving SOCP or SDP is computationally very expensive, which limits their use in practice. Here, we propose an efficient algorithm to solve our objective function in Equation (7).

Taking the derivative with respect to $\mathbf{W}$, and setting the derivative to zero, we have[1]

$$\mathbf{X}\mathbf{X}^T\mathbf{W} - \mathbf{X}\mathbf{Y}^T + \gamma_1\mathbf{D}\mathbf{W} + \gamma_2\tilde{\mathbf{D}}\mathbf{W} = \mathbf{0}, \tag{8}$$

where $\mathbf{D}$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{2||\mathbf{W}^k||_F}\mathbf{I}_k$, $\mathbf{I}_k$ is an identity matrix with size of m_k, $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2||\mathbf{w}^i||_2}$. Thus we have

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\tilde{\mathbf{D}})^{-1}\mathbf{X}\mathbf{Y}^T, \tag{9}$$

where $\mathbf{W}$ can be efficiently obtained by solving the linear equation $(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D} + \gamma_2\tilde{\mathbf{D}})\mathbf{W} = \mathbf{X}\mathbf{Y}^T$, and the matrix inversion that is computationally expensive is not involved.

Note that $\mathbf{D}$ and $\tilde{\mathbf{D}}$ in Equation (9) depend on $\mathbf{W}$ and thus are also unknown variables. We propose an iterative algorithm to solve this problem, which is described in Algorithm 1.

3.3 Analysis of the algorithm

Now we prove that Algorithm 1 converges to the global optimum.

LEMMA 1. *For any matrices* $\mathbf{M}$ *and* $\mathbf{M}_0$ *with the same size, we have* $||\mathbf{M}||_F - \frac{||\mathbf{M}||_F^2}{2||\mathbf{M}_0||_F} \le ||\mathbf{M}_0||_F - \frac{||\mathbf{M}_0||_F^2}{2||\mathbf{M}_0||_F}$.

Proof: Obviously, $-(||\mathbf{M}||_F - ||\mathbf{M}_0||_F)^2 \le \mathbf{M}$, so we have

$$\begin{aligned} &-(||\mathbf{M}||_F - ||\mathbf{M}_0||_F)^2 \le \mathbf{M} \\ \Rightarrow\ &2||\mathbf{M}||_F||\mathbf{M}_0||_F - ||\mathbf{M}||_F^2 \le ||\mathbf{M}_0||_F^2 \\ \Rightarrow\ &||\mathbf{M}||_F - \frac{||\mathbf{M}||_F^2}{2||\mathbf{M}_0||_F} \le ||\mathbf{M}_0||_F - \frac{||\mathbf{M}_0||_F^2}{2||\mathbf{M}_0||_F} \end{aligned}$$

which completes the proof. □

[1]When $\|\mathbf{W}^k\|_F = 0$, the k-th diagonal block of $\mathbf{D}$ can be regularized as $\frac{1}{2\sqrt{\|\mathbf{W}^k\|_F^2 + \varsigma}}\mathbf{I}_k$. Similarly, when $\mathbf{w}^i = 0$, the i-th diagonal element of $\tilde{\mathbf{D}}$ can be regularized as $\frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2 + \varsigma}}$. Then the derived algorithm can be proved to minimize $\sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{y}_i||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\|\mathbf{W}^k\|_F^2 + \varsigma} + \gamma_2 \sum_{i=1}^{d} \sqrt{\|\mathbf{w}^i\|_2^2 + \varsigma}$. It is easy to see that this problem is reduced to problem (6) when $\varsigma \to 0$.

Input: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n] \in \mathbb{R}^{c \times n}$
Initialize $\mathbf{W}_1 \in \mathbb{R}^{d \times c}$, $t = 1$;
while *not converge* **do**
1. Calculate the block diagonal matrix $\mathbf{D}_t$, where the k-th diagonal is $\frac{1}{2||\mathbf{W}_t^k||_F}\mathbf{I}_k$; Calculate the diagonal matrix $\tilde{\mathbf{D}}_t$, where the i-th diagonal element is $\frac{1}{2||\mathbf{w}_t^i||_2}$;
2. $\mathbf{W}_{t+1} = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_t + \gamma_2\tilde{\mathbf{D}}_t)^{-1}\mathbf{X}\mathbf{Y}^T$;
3. $t = t + 1$;
end
Output: $\mathbf{W}_t \in \mathbb{R}^{d \times c}$.

Algorithm 1: Algorithm to solve Equation (7).

THEOREM 1. *Algorithm 1 decreases the objective value in each iteration.*

Proof: In each iteration t, according to Step 2 we have

$$\begin{aligned} &||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 Tr\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1} + \gamma_2 Tr\mathbf{W}_{t+1}^T\tilde{\mathbf{D}}_t\mathbf{W}_{t+1} \\ \le\ &||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 Tr\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t + \gamma_2 Tr\mathbf{W}_t^T\tilde{\mathbf{D}}_t\mathbf{W}_t \\ \Rightarrow\ &||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} \frac{||\mathbf{W}_{t+1}^k||_F^2}{2||\mathbf{W}_t^k||_F} + \gamma_2 \sum_{i=1}^{d} \frac{||\mathbf{w}_{t+1}^i||_2^2}{2||\mathbf{w}_t^i||_2} \\ \le\ &||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} \frac{||\mathbf{W}_t^k||_F^2}{2||\mathbf{W}_t^k||_F} + \gamma_2 \sum_{i=1}^{d} \frac{||\mathbf{w}_t^i||_2^2}{2||\mathbf{w}_t^i||_2}. \end{aligned} \tag{10}$$

Applying Lemma 1 twice to Equation (10), we have the following

$$\begin{aligned} &||\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}_{t+1}^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}_{t+1}^i||_2 \\ \le\ &||\mathbf{W}_t^T\mathbf{X} - \mathbf{Y}||_F^2 + \gamma_1 \sum_{k=1}^{K} ||\mathbf{W}_t^k||_F + \gamma_2 \sum_{i=1}^{d} ||\mathbf{w}_t^i||_2. \end{aligned} \tag{11}$$

Thus, Algorithm 1 decreases the objective value in each iteration. □

Algorithm 1 stops when the following criterion is satisfied:

$$||\mathbf{W}_{t+1} - \mathbf{W}_t||_F / \max\left(||\mathbf{W}_t||_F, 1\right) \le \text{Tol}, \tag{12}$$

where $\text{Tol} = 10^{-4}$ is empirically selected in our experiments.

Upon convergence, $\mathbf{W}_t$, $\mathbf{D}_t$ and $\tilde{\mathbf{D}}_t$ will satisfy Eq. (9). As the problem of solving Eq. (7) is a convex problem, satisfying the Eq. (9) indicates that $\mathbf{W}_t$ is a global optimum solution to Eq. (7). Therefore, Algorithm 1 converges to the global optimum of Eq. (7). Since we have a closed form solution in each iteration, our algorithm converges very fast, which makes our method suitable for not only candidate SNP, but also GWASs.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we evaluate the proposed G-SMuRFS method by applying it to the data from the ADNI cohort, where a wide range of SNPs are examined and selected to predict the response of the MRI imaging phenotypes. The goal is to select a compact set of SNPs while maintaining high predictive power.

4.1 Improved imaging phenotype prediction

We first evaluate the proposed method in predicting the continuous responses of candidate neuroimaging phenotypes. Given two sets of imaging phenotypes, FreeSurfer and VBM, we conduct experiments on each of them separately.

We compare our method against multivariate linear regression, RR and multi-task feature learning (MTFL) (Argyriou *et al.*, 2007)

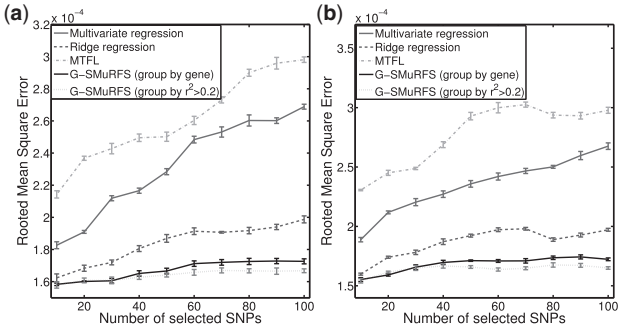

Fig. 5. Performance comparison: The mean and SD of the root mean square errors (RMSEs) obtained from five cross-validation trials in each experiment are plotted, where each error bar indicates ±1 SD. (**a**) FreeSurfer imaging phenotypes; (**b**) VBM imaging phenotypes.

method. The former two are the most widely used methods in statistical learning and medical image analysis. The latter one is a method most related to the proposed method in that it also selects features (SNPs) across tasks; however, it only uses $\ell_{2,1}$-norm regularization whereby group information is not taken into account. Therefore, MTFL method can be seen as a special case of the proposed method by setting $\gamma_1 = 0$ in Equation (7).

We group SNPs using two methods: (i) SNPs annotated with the same gene are grouped together; (ii) SNPs within the same LD block are grouped together, where $r^2 \geq 0.2$ is used in this work. For each test case, we conduct standard 5-fold cross-validation and report the average results. For each of the five trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right\}$ for RR, MTFL method and our method. For each trial, from the learned coefficient matrix we sum the absolute values of the coefficients of a single SNP over all the tasks as the SNP weight, from which we pick up the top $\{10, 20, \ldots, 100\}$ SNPs to predict the regression responses for the test data. The performance of each trial is assessed by RMSE, a widely used measurement for regression analysis. For each experiment, the mean and standard deviation (SD) of the RMSEs obtained from the five trials are reported in Figure 5, where each error bar indicates ±1 SD. Detailed RMSE results of each fold in cross-validation are available in the Supplementary Material at http://ranger.uta.edu/%7eheng/imaging-genetics/.

The proposed G-SMuRFS methods consistently outperform three competing methods in both FreeSurfer and VBM cases (Fig. 5), while the cross-validation trials in each experiment perform very similarly to one another (see the error bars in Fig. 5). For a formal comparison, *t*-test is performed and the resulting *P*-values are reported in Table 2, from which we can see that our methods are significantly better than three competing methods. Moreover, the predictive performances of our methods are considerably stable, whereas those of the other methods are sensitive to experimental conditions. These results clearly demonstrate the advantage of the proposed G-SMuRFS method in predicting phenotypic responses.

A more careful observation shows that the regression performance of our method when using $r^2 > 0.2$ to group SNPs is better than that of our method when grouping SNPs by genes. While gene is the most natural way to group SNPs, different segments within the same gene may have different functions (e.g. bases for different isoforms) and mixing them together may perturb the prediction. Grouping by LD blocks using r^2 yields more homogeneous groups and has a potential to boost the prediction power.

Table 2. The results (*P*-values) of *t*-tests for performance comparison between our methods and three competing methods

	FreeSurfer biomarkers		VBM biomarkers	
	Group by gene	Group by $r^2 > 0.2$	Group by gene	Group by $r^2 > 0.2$
MLR	7.08×10^{-5}	3.13×10^{-5}	9.28×10^{-6}	3.96×10^{-6}
RR	1.31×10^{-2}	2.21×10^{-3}	2.12×10^{-2}	1.85×10^{-3}
MTFL	6.57×10^{-7}	2.41×10^{-7}	5.82×10^{-7}	2.63×10^{-7}

Figure 6 shows heat maps of prediction errors on each QT. While all these QTs are AD-relevant, Figure 6 indicates that they are affected in different degrees by genetic factors. QTs that are better predicted by SNPs include GM density measures of the parahippocampal gyrus and frontal region in VBM analyses and thickness measures of the frontal region, lateral temporal region and precuneus in FreeSurfer analyses. The VBM and FreeSurfer measures of a certain region yield similar results in some cases (e.g. frontal region), but may provide different information in other cases (e.g. parahippocampal gyrus). Thus, performing both VBM and FreeSurfer analyses can help identify useful imaging phenotypes and guide further investigation to better elucidate the underlying disease mechanism, from gene, to brain structure and function, and to symptoms.

4.2 Genetic marker selection

Shown in Figure 7 are the regression coefficients for top 10 selected SNPs. First, these SNPs are either AlzGene candidates or proximal to the candidates; however, little is known about their underlying mechanisms in relation to AD. The results shown in Figure 7 can help identify relevant QTs for each SNP and has a potential to gain biological insights from gene to brain to symptoms. Second, as expected, the *APOE* SNP rs429358 shows the strongest association with all QTs in each experiment; and the hippocampal measures exhibit the strongest association with the *APOE* SNP. Clearly, the proposed approach is able to identify the most important AD genetic risk factor via imaging QTs as well as the best-known neurodegenerative marker. Third, besides confirming the prior findings, our method also yielded new discoveries such as the associations between *APOE* and other eminent AD markers including entorhinal cortex and parahippocampal gyrus. These associations were not identified in our prior massive univariate analyses on the same data (Shen *et al.*, 2010), indicating that the proposed multilocus method has increased power to discover interesting imaging QTs. In sum, the above evidence demonstrates not only the effectiveness of the proposed method, but also the strength of using imaging QTs in genetic association study.

Quite a few SNPs from the *SORCS1* gene are selected as the top 10 hits in each experiment; however, the large size of the gene (Fig. 1) may play a role. Figure 8 shows an LD plot with location maps for a group of 46 *SORCS1* SNPs, where two top hits (red spikes) are highlighted for each of the FreeSurfer and VBM experiments. Although *SORCS1* has been associated with

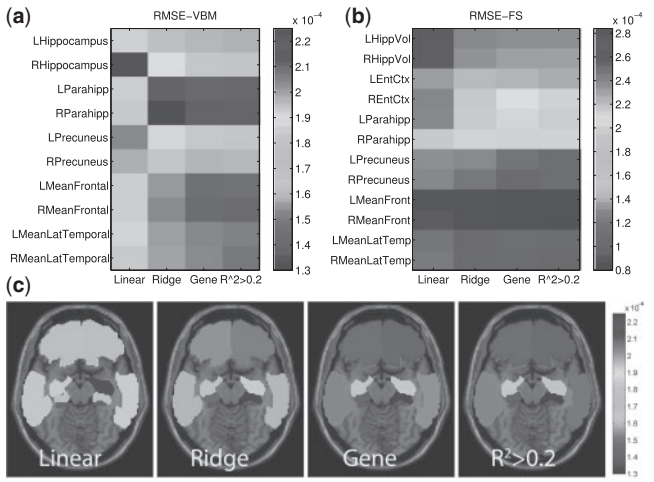

Fig. 6. (**a** and **b**) Show the heat maps of RMSEs for predicting VBM (a) and FreeSurfer (b) measures using LR, RR, our G-SMuRFS method with SNPs grouped by gene and G-SMuRFS with SNPs grouped by $r^2 > 0.2$, where top 10 SNPs were used in our G-SMuRFS methods. In (**c**), RMSEs for predicting VBM measures using four methods are mapped onto the brain volume.

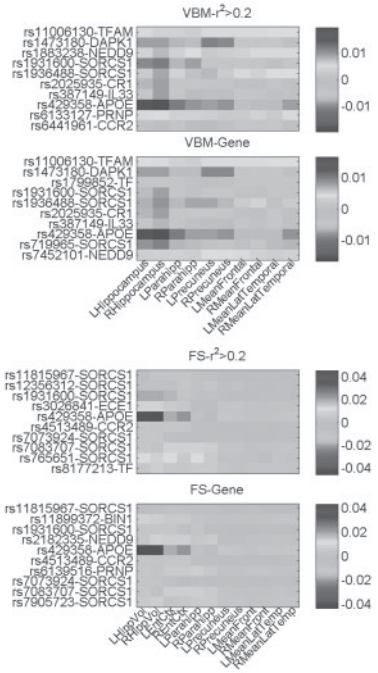

Fig. 7. Regression coefficients are visualized for top 10 selected SNPs in each of the four experiments (from top to bottom): (i) group by $r^2 > 0.2$, regression on VBM measures; (ii) group by gene, regression on VBM measures; (iii) group by $r^2 > 0.2$, regression on FreeSurfer measures; and (iv) group by gene, regression on FreeSurfer measures.

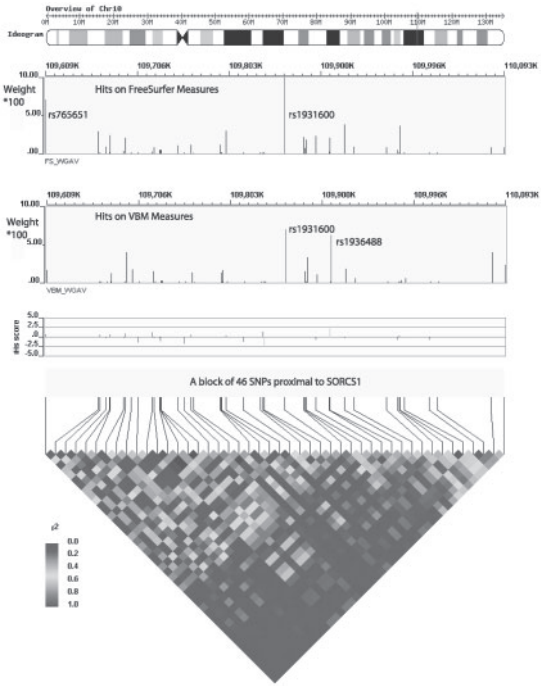

Fig. 8. Pair-wise LD in a group of 46 SNPs proximal to SORCS1. Numerical values r^2 of the LD maps are determined by Haploview and visualized with WGAViewer. The top panel is the ideogram of the chromosome and the vertical red line represents the relative location of the locus of interest. In the second panel, regression coefficients*100 is plotted for each SNP for the FreeSurfer data, where two top hits rs765651 and rs1931600 are labeled with red lines. In the third panel, regression coefficients*100 is plotted for each SNP for the VBM data, where two top hits rs1931600 and rs1936488 are labeled with red lines. The fourth panel shows the recent selection score (Voight *et al.*, 2006). The bottom figure demonstrates the LD pattern among 46 SNPs.

diabetes and AD (Lane *et al.*, 2010), the top ranked *SORCS1* SNPs in Figure 7 have not been reported in prior association studies. Thus, this gene together with its SNPs warrants further investigation in independent cohorts. Due to the nature of our method, an epistasis analysis on these top hits would be appropriate for investigation in future studies.

5 CONCLUSIONS

In this article, we have proposed a novel G-SMuRFS method to perform both regression analysis for predicting continuous responses of brain imaging measures and selecting relevant SNPs in an MCI/AD study. Different from traditional regression methods that ignore the interrelated structures within genotyping and imaging data, our method studies the associations between SNPs and imaging phenotypes within a single regression framework and shared common subspace. Through enforcing a new form of regularization using $G_{2,1}$-norm that takes into account both group-level structural information inside SNP data and sparsity among SNP groups, our learning model is able to exploit additional information to achieve both enhanced predictive performance and improved feature (SNP) selection capability. Besides, $\ell_{2,1}$-norm regularization is used in our model to jointly select SNPs relevant to important imaging phenotypes. An efficient algorithm to solve the proposed objective is presented with rigorous proof of its correctness and convergence. Our experiments using the SNP and MRI data from the ADNI cohort yielded the following promising results: (i) the prediction performance of G-SMuRFS method was consistently better than conventional multi-variate linear regression and ridge regression; (ii) a compact set of SNP predictors were identified in each test case, warranting further investigation in independent cohorts for confirmation; and (3) these selected SNPs could predict the responses of multiple imaging phenotypes at the same time and had a potential to serve as useful genetic risk factors for AD. These promising results were consistent with our theoretical foundation and in accordance with some prior studies, which demonstrated the effectiveness of the proposed method.

One important future direction of this work could be to explore the possibility of simultaneously employing multiple SNP grouping schemes or more generally adopting a pre-defined network/pathway strategy and see whether these approaches can further improve the prediction performance. Other potential future directions include (i) application of G-SMuRFS method to additional imaging phenotypes (e.g. PET, fMRI data); and (ii) building a principled sparse learning framework to reveal complex relationships among multiple data sources available in the ADNI database, including genetic, cerebrospinal fluid, plasma, imaging and cognitive datasets to study AD at a system biology level.

Funding: This research is supported by National Science Foundation Grants CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965 at UTA; and by National Science Foundation Grant IIS-1117335, National Institutes of Health Grants UL1 RR025761, U01 AG024904, NIA RC2 AG036535, NIA R01 AG19771, and NIA P30 AG10133-18S1 at IU. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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 California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

Conflict of Interest: none declared.

REFERENCES

Argyriou,A. *et al.* (2007) Multi-task feature learning. *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, pp. 41–48.

Ashburner,J. and Friston,K. (2000) Voxel-based morphometry–the methods. *Neuroimage*, **11**, 805–821.

Ballard,D.H. *et al.* (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.*, **34**, 201–212.

Barrett,J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Bertram,L. *et al.* (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.*, **39**, 17–23.

Bralten,J. *et al.* (2011) Association of the Alzheimer's gene SORL1 with hippocampal volume in young, healthy adults. *Am. J. Psychiatry*, **168**, 1083–1089.

Braskie,M.N. *et al.* (2011) Neuroimaging measures as endophenotypes in Alzheimer's disease. *Int. J. Alzheimers Dis.*, **2011**, 490140.

Fischl,B. *et al.* (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, **33**, 341–355.

Glahn,D. *et al.* (2007) Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. *Hum. Brain Mapp.*, **28**, 488–501.

Hibar,D.P. *et al.* (2011) Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*, **56**, 1875–1891.

Hinrichs,C. *et al.* (2011) Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage*, **55**, 574–589.

Lane,R.F. *et al.* (2010) Diabetes-associated SorCS1 regulates Alzheimer's amyloid-beta metabolism: evidence for involvement of SorL1 and the retromer complex. *J. Neurosci.*, **30**, 13110–13115.

Lee,S. *et al.* (2010) Adaptive Multi-Task Lasso: with application to eQTL detection. *Adv. Neural Informat. Process. Syst.*, 1306–1314.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Obozinski,G. *et al.* (2006) Multi-task feature selection. *Technical Report*, Department of Statistics, University of California, Berkeley.

Potkin,S.G. *et al.* (2009) Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cogn. Neuropsychiatry*, **14**, 391–418.

Puniyani,K. *et al.* (2010) Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, **26**, i208.

Saykin,A.J. *et al.* (2010) Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement.*, **6**, 265–273.

Shen,L. *et al.* (2010) 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*, **53**, 1051–1063.

Stein,J.L. *et al.* (2010) Voxelwise genome-wide association study (vGWAS). *Neuroimage*, **53**, 1160–1174.

Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B.*, **58**, 267–288.

Voight,B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.

Walhovd,K. *et al*. (2010) Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol. Aging*, **31**, 1107–1121.

Weiner,M.W. *et al.* (2010) The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*., **6**, 202–211.e7.

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Yip,W. K. and Lange,C. (2011) Quantitative trait prediction based on genetic marker-array data, a simulation study. *Bioinformatics*, **27**, 745–748.

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

Zhan,H. *et al.* (2011) A stochastic expectation and maximization algorithm for detecting quantitative trait-associated genes. *Bioinformatics*, **27**, 63–69.

Zhang,D. *et al.* (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, **55**, 856–867.