BIOINFORMATICS

Vol. 30 ECCB 2014, pages i564–i571
doi:10.1093/bioinformatics/btu465

Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm

Jingwen Yan[1,2], Lei Du[2], Sungeun Kim[2], Shannon L. Risacher[2], Heng Huang[3], Jason H. Moore[4], Andrew J. Saykin[2], Li Shen[2,*] and for the Alzheimer's Disease Neuroimaging Initiative

[1]BioHealth, Indiana University School of Informatics & Computing, [2]Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA, [3]Computer Science & Engineering, The University of Texas at Arlington, TX 76019, USA and [4]Genetics, Community & Family Medicine, Dartmouth Medical School, Lebanon, NH 03756, USA

ABSTRACT

Motivation: Imaging genetics is an emerging field that studies the influence of genetic variation on brain structure and function. The major task is to examine the association between genetic markers such as single-nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. The complexity of these datasets has presented critical bioinformatics challenges that require new enabling tools. Sparse canonical correlation analysis (SCCA) is a bi-multivariate technique used in imaging genetics to identify complex multi-SNP–multi-QT associations. However, most of the existing SCCA algorithms are designed using the soft thresholding method, which assumes that the input features are independent from one another. This assumption clearly does not hold for the imaging genetic data. In this article, we propose a new knowledge-guided SCCA algorithm (KG-SCCA) to overcome this limitation as well as improve learning results by incorporating valuable prior knowledge.

Results: The proposed KG-SCCA method is able to model two types of prior knowledge: one as a group structure (e.g. linkage disequilibrium blocks among SNPs) and the other as a network structure (e.g. gene co-expression network among brain regions). The new model incorporates these prior structures by introducing new regularization terms to encourage weight similarity between grouped or connected features. A new algorithm is designed to solve the KG-SCCA model without imposing the independence constraint on the input features. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, using an Alzheimer's disease (AD) cohort, we examine the imaging genetic associations between all SNPs in the *APOE* gene (i.e. top AD gene) and amyloid deposition measures among cortical regions (i.e. a major AD hallmark). In comparison with a widely used SCCA implementation, our KG-SCCA algorithm produces not only improved cross-validation performances but also biologically meaningful results.

Availability: Software is freely available on request.

Contact: shenli@iu.edu

1 INTRODUCTION

Brain imaging genetics is an emerging field that studies the influence of genetic variation on brain structure and function. Its major task is to examine the association between genetic markers such as single-nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from multimodal neuroimaging data (e.g. anatomical, functional and molecular imaging scans). Given the well-known importance of gene and imaging phenotype in brain function, bridging these two factors and exploring their connections would lead to a better mechanistic understanding of normal or disordered brain functions. The complexity of these data, however, has presented critical bioinformatics challenges requiring new enabling tools. Early studies in imaging genetics typically focused on pairwise univariate analysis (Shen *et al.*, 2010). Many recent studies turned to regression analysis for exploring the joint effect of multiple SNPs on single or few QTs (Hibar *et al.*, 2011) and bi-multivariate analyses for revealing complex multi-SNPs–multi-QTs associations (Chi *et al.*, 2013; Lin *et al.*, 2014; Vounou *et al.*, 2010; Wan *et al.*, 2011).

Canonical correlation analysis (CCA), a bi-multivariate method, has been applied to imaging genetics applications. It aims to find the best linear transformation for imaging and genetics features so that the highest correlation between imaging and genetic components can be achieved. Based on the assumption that a real imaging genetic signal typically involves a small number of SNPs and QTs, sparse canonical correlation analysis (SCCA) has also been applied in several imaging genetic studies by imposing the Lasso regularization term to yield sparse results (Chi *et al.*, 2013; Lin *et al.*, 2014; Wan *et al.*, 2011). However, most existing SCCA algorithms are designed using the soft thresholding technique, which assumes that the input features are independent from one another (Tibshirani, 1996). This assumption clearly does not hold for the imaging genetic data [e.g. the existence of the structural and functional networks in the brain and the linkage disequilibrium (LD) blocks in the genome]. Directly ignoring the covariance structure in the data will inevitably limit the capability of yielding optimal results.

In this article, we propose a new knowledge-guided SCCA algorithm (KG-SCCA) to overcome this limitation as well as to aim for improving learning results by incorporating valuable prior knowledge. The proposed KG-SCCA method is able to model two types of prior knowledge: one as a group structure (e.g. LD blocks among SNPs) and the other as a network structure (e.g. gene co-expression network among brain regions). The new model incorporates these prior structures by introducing new regularization terms to encourage similarity between grouped or connected features. A new algorithm is designed to solve the KG-SCCA model without imposing the independence

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

constraint on the input features. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, using an Alzheimer's disease (AD) cohort, we examine the imaging genetic associations between all SNPs in the *APOE* gene (i.e. top AD gene) and amyloid deposition measures among cortical regions (i.e. a major AD hallmark). In comparison with a widely used SCCA implementation in the PMA software package (http://cran.r-project.org/web/packages/PMA/) (Witten *et al.*, 2009), our KG-SCCA algorithm produces improved cross-validation performances as well as biologically meaningful results.

2 MATERIALS AND DATA SOURCES

To demonstrate the proposed KG-SCCA algorithm, we apply it to an amyloid imaging genetic analysis in the study of AD. Deposition of amyloid-β in the cerebral cortex is a major hallmark in AD pathogenesis. Our prior studies (Ramanan *et al.*, 2014; Swaminathan *et al.*, 2012) performed univariate genetic association analyses of amyloid measures in a few candidate cortical regions of interest (ROIs), and identified several promising hits including rs429358 in *APOE*, rs509208 in *BCHE* and rs7551288 in *DHCR24*. In this work, using the proposed KG-SCCA algorithm, we perform a bi-multivariate analysis to examine the association between all the available SNPs (58 in total) in the *APOE* gene (i.e. the top genetic risk factor for late onset AD) and 78 ROIs across the entire cortex. We use two types of prior knowledge in this analysis: (i) a group structure is imposed to the SNP data using the LD block information (Fig. 4), and (ii) a network structure is imposed to the amyloid imaging data by computing an amyloid pathway-based gene co-expression network in the brain using Allen Human Brain Atlas (AHBA; Zeng *et al.*, 2012). Below, we first describe our amyloid imaging and genotyping data, and then discuss our method for creating the amyloid pathway-based gene co-expression network in the brain.

2.1 Imaging and genotyping data

The proposed algorithm, KG-SCCA, was empirically evaluated using the amyloid imaging and genotyping data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). One goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org. Preprocessed [18F]Florbetapir PET scans (i.e. amyloid imaging data) were downloaded from LONI (adni.loni.usc.edu). Before downloading, images were averaged, aligned to a standard space, resampled to a standard image and voxel size, smoothed to a uniform resolution and normalized to a cerebellar gray matter reference region resulting in standardized uptake value ratio images as previously described (Jagust *et al.*, 2010). After downloading, the images were aligned to each participant's same visit MRI scan and normalized to the Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm voxels using parameters from the MRI segmentation. ROI level amyloid measurements were further extracted based on the MarsBaR AAL atlas. Genotype data of both ADNI-1 and ADNI-2/GO phases were also obtained from LONI (adni.loni.usc.edu). All the *APOE* SNPs were extracted based on the quality controlled and imputed data combining two phases together. Only SNPs available in Illumina 610Quad and/or OmniExpress arrays were included in the analysis. As a result, we had 58 SNPs located within 10 LD blocks (Fig. 4) computed using HaploView (Barrett, 2009). A total of 568 non-Hispanic Caucasian participants with both complete amyloid measurements and *APOE* SNPs were studied, including 28 AD, 343 MCI and 196 healthy control (HC) subjects (Table 1). Using the regression weights derived from the HC participants, amyloid and SNP measures were preadjusted for removing the effects of the baseline age, gender, education and handedness.

Table 1. Participant characteristics

Subjects	AD	MCI	HC
Number	28	343	196
Gender (M/F)	18/10	203/140	102/94
Handedness (R/L)	23/5	309/34	178/18
Age (mean ± std)	75.23 ± 10.66	71.92 ± 7.47	74.77 ± 5.39
Education (mean ± std)	15.61 ± 2.74	15.99 ± 2.75	16.46 ± 2.65

2.2 Amyloid pathway-based gene co-expression network in the brain

Because we examine cortical amyloid deposition in relation to genetic variation, we hypothesize that amyloid pathway-based gene co-expression profiles among cortical ROIs may provide valuable information in search for *APOE*-related amyloid distribution pattern in the cortex. Thus, we used the brain transcriptome data from the AHBA (Zeng *et al.*, 2012), coupled with 15 candidate genes from amyloid pathways studied in (Swaminathan *et al.*, 2012), to create such a brain network.

Gene expression profiles across the whole human brain were downloaded from Allen Institute for Brain Science. One of their goals is to advance the research and knowledge about neurobiological conditions, with extensive mapping of whole-genome gene expression throughout the brain. Among various organisms, AHBA is one of the projects seeking to combine the genomics with the neuroanatomy to better understand the connection between genes and brain functioning. Gene expression profiles in eight health human brains have been released, including two full brains and six right hemispheres. Details can be found in www.brain-map.org.

Brain-wide expression data of all 15 amyloid-related candidate genes, reported in (Swaminathan *et al.*, 2012), were extracted from AHBA to construct the brain network. Because an early report indicated that individuals share as much as 95% gene expression profile (Zeng *et al.*, 2012), in this study, we only included two full brains (H0351-2201 and H0351-2002) to construct the co-expression network. First all the brain samples (~900) in AHBA were mapped to MarSBAR AAL atlas, which included 116 brain ROIs. According to Ramanan *et al.* (2014), cortical ROIs are typically believed to hold the amyloid

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

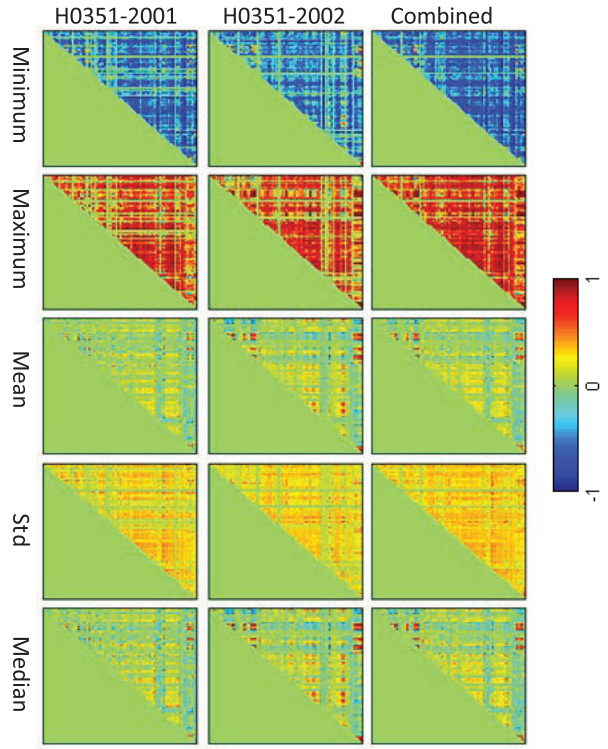


Fig. 1. Amyloid pathway-based gene co-expression networks among 78 AAL cortical ROIs constructed from AHBA using different statistics (see different rows) for two individuals and their combination

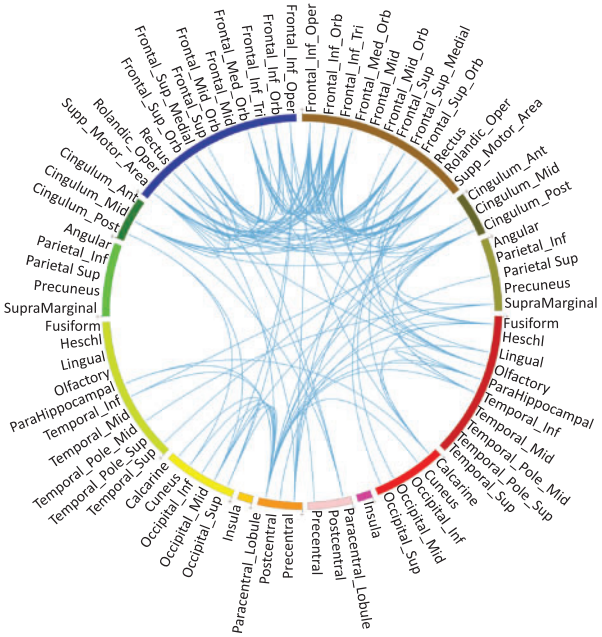


Fig. 2. Network visualization by thresholding the connectivity matrix shown in the lower right corner of Figure 1, where edges correspond to matrix entries with values ≥ 0.5 or ≤ -0.5. The circle is symmetric (left measures on left and right measures on right), from top to bottom are frontal lobe, cingulate, parietal lobe, temporal lobe, occipital lobe, insula and sensory–motor cortex

signals, whereas other ROIs hold similar amyloid measures across individuals. Thus, 39 pairs of bilateral cortical ROIs (78 in total), from frontal lobe, cingulate, parietal lobe, temporal lobe, occipital lobe, insula and sensory–motor cortex, were included in our analysis. Correlation among ~900 brain locations was first calculated based on the gene expression profile of 15 amyloid candidate genes. Due to many-to-one mapping from the brain locations to AAL ROIs, for each ROI, there are more than one connections, represented by correlations between two brain locations. Therefore, we calculated ROI-level correlations of two individuals in five ways: minimum, maximum, mean, standard deviation and median. In addition, the ROI correlation structure based on the combination of both individuals was also generated in the same way for comparison (Fig. 1). Clearly, for all five statistics, the pattern remains highly consistent across individuals and their combination. For simplicity, in the subsequent analysis, we adopt the brain connectivity matrix generated from the combination sample using the median statistics (i.e. the panel in the lower right corner of Fig. 1). Figure 2 shows a network visualization of this matrix, where edges correspond to matrix entries with values ≥ 0.5 or ≤ -0.5.

3 METHODS

Now we present our KG-SCCA algorithm. We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M}=(m_{ij})$, we denote its i-th row and j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. Let $\mathbf{X}=\{x_1, ..., x_n\} \subseteq \Re^p$ be the genotype data (SNP) and $\mathbf{Y}=\{y_1, ..., y_n\} \subseteq \Re^q$ be the imaging QT data, where n is the number of participants, p and q are the numbers of SNPs and QTs, respectively.

CCA seeks linear transformations of variables $\mathbf{X}$ and $\mathbf{Y}$ to achieve the maximal correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$, which can be formulated as:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad s.t.\ \mathbf{u}^T\mathbf{X}^T\mathbf{Xu}=1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Yv}=1 \tag{1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical loadings or weights, reflecting the significance of each feature in the identified canonical correlation.

Similar to many machine learning algorithms, overfitting could arise in CCA when the features outnumber the participants. In addition, the CCA outcome could spread non-trivial effects across all the features rather than only a few significant ones, making the results difficult to interpret. To address these issues, SCCA was proposed in (Witten *et al.*, 2009) by introducing penalty terms, $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$, to regularize the weights, as shown in Equation (2).

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \\ &s.t. \quad ||\mathbf{Xu}||_2^2=1, ||\mathbf{Yv}||_2^2=1, P_1(\mathbf{u}) \leq c_1, P_2(\mathbf{v}) \leq c_2 \end{aligned} \tag{2}$$

Here the objective function is bilinear in $\mathbf{u}$ and $\mathbf{v}$: when $\mathbf{u}$ is fixed, it is linear in $\mathbf{v}$ and vice versa. But due to the L_2 equality, with $\mathbf{u}$ or $\mathbf{v}$ fixed, the constraints are not convex. This can be solved by reformulating the L_2 equality into inequality as $||\mathbf{Xu}||_2^2 \leq 1$ and $||\mathbf{Yv}||_2^2 \leq 1$. For easy computation, Equation (2) is commonly rewritten in its Lagrangian form.

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} - \frac{\gamma_1}{2}||\mathbf{Xu}||_2^2 - \frac{\gamma_2}{2}||\mathbf{Yv}||_2^2 - \beta_1 P_1(\mathbf{u}) - \beta_2 P_2(\mathbf{v}) \tag{3}$$

Witten *et al.* (2009) and Witten and Tibshirani (2009) explored two penalty forms, L_1 penalty and the chain structured fused Lasso penalty. L_1 penalty imposes sparsity on both $\mathbf{u}$ and $\mathbf{v}$ and assumes that each canonical correlation involves only a few features from $\mathbf{X}$ and $\mathbf{Y}$. The fused

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Lasso penalty promotes the smoothness of weight vectors and encourages neighboring features to be selected together. To incorporate other structures, group- and network-guided penalties were introduced (Chen and Liu, 2012; Chen *et al.*, 2013). As mentioned earlier, most of these methods were designed using the soft thresholding technique, which was first proposed to solve Lasso problem when the features were independent from each other (Tibshirani, 1996). This condition does not hold in imaging genetics data. Thus, direct application of those methods into imaging genetics studies limits the capability of yielding optimal solutions. Below, we first present our KG-SCCA model and then present an effective KG-SCCA algorithm without using the soft thresholding strategy.

Brain has been studied as a complicated network. The SNP data have structures like LD blocks. Given these prior knowledge, we propose the following KG-SCCA model by introducing two penalty terms for genetic loadings $\mathbf{u}$ and imaging loading $\mathbf{v}$, respectively.

$$\begin{aligned} P_1 = ||\mathbf{u}||_G &= \beta_1 \sum_{k_1=1}^{K_1} \sqrt{\sum_{i \in \pi_{k_1}} u_i^2} + \theta_1 ||\mathbf{u}||_1 \\ &= \beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 + \theta_1 ||\mathbf{u}||_1, \end{aligned} \tag{4}$$

$$\begin{aligned} P_2 = ||\mathbf{v}||_N &= \beta_2 \sum_{\substack{(i,j)\in E \\ i<j}} \tau(w_{ij}) ||v_i - sign(w_{ij}) v_j||_2^2 + \theta_2 ||\mathbf{v}||_1 \\ &= \beta_2 ||\mathbf{Cv}||_2^2 + \theta_2 ||\mathbf{v}||_1. \end{aligned}$$

In penalty $P_1(\mathbf{u})$, SNPs are partitioned into K_1 groups $\Pi_1 = \{\pi_{k_1}\}_{k_1=1}^{K_1}$, such that $\{u_i\}_{i=1}^{m_{k_1}} \in \pi_{k_1}$, and m_{k_1} is the number of SNPs in π_{k_1}. While the group term $\beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2$ helps select all the SNPs in relevant LD blocks, L_1 penalty manages to suppress those non-signals within selected LD blocks. The $P_1(\mathbf{u})$ penalty is essentially the group Lasso penalty applied to the CCA framework.

Penalty $P_2(\mathbf{v})$ applies the network-guided constraint to encourage the joint selection of 'connected' features (i.e. their connectivity matrix entry having a high weight) as well as uses L_1 to impose global sparsity. E is the set of all possible imaging QT pairs and $|E|$ is the total number of QT pairs. $\mathbf{C} \in \Re^{|\mathbf{E}|\times q}$ is defined as follows. The row of $\mathbf{C}$ is indexed by all pairs $(i,j) \in \{(i,j)|i \in \{1,...,q\}, j \in \{1,...q\}, i<j\}$, $\mathbf{C}_{(i,j),i} = w_{ij}$ and $\mathbf{C}_{(i,j),j} = sign(w_{ij}) w_{ij}$. $\tau(w_{ij})$ provide the fusion effect that promotes similarity between v_i and v_j of related features. In this article, we use $\tau(w_{ij}) = w_{ij}^2$. With $sign(w_{ij})$ we can have positively related features being pulled together and on the other hand the negatively related features being fused with opposite direction. Thus, for strongly connected features with a large fusion effect, they tend to be jointly selected or jointly not selected.

In this work, as mentioned earlier, we formed the group structure for the SNP data by partitioning them using LD blocks generated by HaploView (Barrett, 2009). We formed the network structure for the amyloid imaging data by constructing amyloid pathway-based gene co-expression network using AHBA. Because the model could be easily extended to estimate multiple canonical variables, we only focus on creating the first pair of canonical variables in this article.

Algorithm 1 Knowledge-guided SCCA (KG-SCCA)

Require:

$\mathbf{X} = \{x_1, ..., x_n\}$, $\mathbf{Y} = \{y_1, ..., y_n\}$, group and network structures

Ensure:

Canonical vectors $\mathbf{u}$ and $\mathbf{v}$.

$t = 1$, Initialize $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$;
while not converge **do**
  Calculate $\mathbf{B}_{1_t} = \frac{1}{\gamma_1}\mathbf{Y}\mathbf{v}_t$
  Calculate the block diagonal matrix $\mathbf{D}_{1_t}$ and $\mathbf{D}_{2_t}$;
  $\mathbf{u}_{t+1} = (\mathbf{X}^T\mathbf{X} + \frac{\beta_1}{\gamma_1}\mathbf{D}_{1_t} + \frac{\theta_1}{\gamma_1}\mathbf{D}_{2_t})^{-1}\mathbf{X}^T\mathbf{B}_{1_t}$;
  Scale $\mathbf{u}_{t+1}$ so that $\mathbf{u}_{t+1}^T\mathbf{X}^T\mathbf{X}\mathbf{u}_{t+1} = 1$;
  Calculate $\mathbf{B}_{2_t} = \frac{1}{\gamma_2}\mathbf{X}\mathbf{u}_{t+1}$;
  Calculate the block diagonal matrix $\mathbf{D}_{4_t}$;
  $\mathbf{v}_{t+1} = (\mathbf{Y}^T\mathbf{Y} + \frac{\beta_2}{\gamma_2}\mathbf{D}_3 + \frac{\theta_2}{\gamma_2}\mathbf{D}_{4_t})^{-1}\mathbf{Y}^T\mathbf{B}_{2_t}$;
  Scale $\mathbf{v}_{t+1}$ so that $\mathbf{v}_{t+1}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v}_{t+1} = 1$;
  $t = t + 1$.
end while

We now present our algorithm to solve this model without using soft thresholding approach. By fixing $\mathbf{u}$ and $\mathbf{v}$, respectively, we will have two convex problems shown in Equation (5).

$$\begin{aligned} &\max_{\mathbf{u}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}||\mathbf{Xu}||_2^2 - \beta_1 \sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 - \theta_1 ||\mathbf{u}||_1 \\ &\max_{\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_2}{2}||\mathbf{Yv}||_2^2 - \frac{\beta_2}{2}||\mathbf{Cv}||_2^2 - \theta_2 ||\mathbf{v}||_1 \end{aligned} \tag{5}$$

Let $\mathbf{B}_1 = \frac{1}{\gamma_1}\mathbf{Y}\mathbf{v}$ and $\mathbf{B}_2 = \frac{1}{\gamma_2}\mathbf{X}\mathbf{u}$, the above problems can be reformulated to Equation (6):

$$\begin{aligned} &\min_{\mathbf{u}} \frac{1}{2}||\mathbf{Xu} - \mathbf{B}_1||_2^2 + \frac{\beta_1}{\gamma_1}\sum_{k_1=1}^{K_1} ||\mathbf{u}^{k_1}||_2 + \frac{\theta_1}{\gamma_1}||\mathbf{u}||_1 \\ &\min_{\mathbf{v}} \frac{1}{2}||\mathbf{Yv} - \mathbf{B}_2||_2^2 + \frac{\beta_2}{2\gamma_2}||\mathbf{Cv}||_2^2 + \frac{\theta_2}{\gamma_2}||\mathbf{v}||_1 \end{aligned} \tag{6}$$

Here, while $\mathbf{u}$ can be solved by the G-SMuRFS method proposed in (Wang *et al.*, 2012), optimization of $\mathbf{v}$ can be achieved by the network-guided $L_{2,1}$ regression method proposed in (Yan *et al.*, 2013). In both solutions, a smooth approximation has been estimated for group $L_{2,1}$ and L_1 terms by including an extremely small value. The solution for $\mathbf{u}$ and $\mathbf{v}$ in each iteration step is as follows:

$$\begin{aligned} \mathbf{u} &= (\mathbf{X}^T\mathbf{X} + \frac{\beta_1}{\gamma_1}\mathbf{D}_1 + \frac{\theta_1}{\gamma_1}\mathbf{D}_2)^{-1}\mathbf{X}^T\mathbf{B}_1, \\ \mathbf{v} &= (\mathbf{Y}^T\mathbf{Y} + \frac{\beta_2}{\gamma_2}\mathbf{D}_3 + \frac{\theta_2}{\gamma_2}\mathbf{D_4})^{-1}\mathbf{Y}^T\mathbf{B}_2, \end{aligned} \tag{7}$$

where $\mathbf{D_1}$ is a block diagonal matrix with the k-th diagonal block as $\frac{1}{\|\mathbf{u}^k\|_F}\mathbf{I}_k$; $\mathbf{I}_k$ is an identity matrix with size of m_k; m_k is the total feature number in group k; $\mathbf{D}_2$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{u}^i\|_2}$; $\mathbf{D}_3 = \mathbf{C}^T\mathbf{C}$ is a matrix in which each row integrates all the neighboring relationships (e.g. for the i-th row, it is the sum of all the rows in α whose i-th element is not zero); and $\mathbf{D_4}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{v}^i\|_2}$. Algorithm 1 summarizes the KG-SCCA optimization procedure. Further details on how to solve for two objectives in Equation (6) are available in (Wang *et al.*, 2012) and (Yan *et al.*, 2013), respectively.

In Algorithm 1, six parameters $\gamma_1, \gamma_2, \beta_1, \beta_2, \theta_1, \theta_2$ need to be tuned to control the global sparsity as well as structured group or network constraints. Chen and Liu (2012) studied a similar problem using a different method, and found that their results were insensitive to γ_1, γ_2 settings. Following their observation, we set γ_1 and γ_2 to 1 for simplicity. Nested cross-validation can be used for parameter selection but will be extremely time-consuming for the remaining four parameters. Thus, we followed the strategy proposed in (Lin *et al.*, 2014): parameters β_1, β_2 controlling structural constraints were first tuned without considering sparsity constraints. Then based on the obtained optimal β_1, β_2, another nested cross-validation was performed to acquire the optimal θ_1, θ_2.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

We performed comparative studies between the proposed KG-SCCA algorithm and a widely used SCCA implementation

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

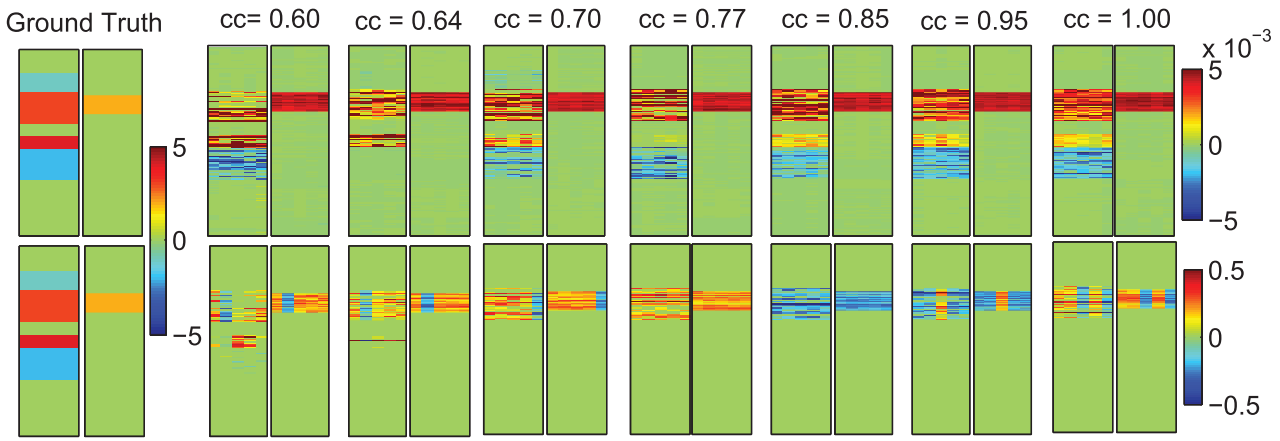


Fig. 3. Five-fold trained weights of **u** and **v**. Ground truth of **u** and **v** are shown in the most left two panels. KG-SCCA results (top row) and PMA results (bottom row) are shown in the remaining panels, corresponding to true correlation coefficients (CCs) ranging from 0.6 to 1.0. For each panel pair, the five estimated **u** values are shown on the left panel, and the five estimated **v** values are shown on the right panel

Table 2. Five-fold cross-validation performance on synthetic data: mean ± std is shown for estimated correlation coefficients and AUC of the test data using the trained model

True	Correlation coefficients (CC)			AUC				
CC	KG-SCCA	PMA	*P*	KG-SCCA:u	PMA:u	*P*	KG-SCCA:v	PMA:v
0.60	0.56 ± 0.12	0.31 ± 0.14	2.19E-03	0.83 ± 0.08	0.64 ± 0.02	3.36E-03	1.0 ± 0.00	1.0 ± 0.00
0.64	0.56 ± 0.1	0.51 ± 0.12	2.32E-02	0.96 ± 0.04	0.65 ± 0.01	2.20E-05	1.0 ± 0.00	1.0 ± 0.00
0.70	0.64 ± 0.1	0.53 ± 0.1	1.27E-05	0.99 ± 0.01	0.62 ± 0.	6.21E-08	1.0 ± 0.00	1.0 ± 0.00
0.77	0.7 ± 0.14	0.6 ± 0.14	6.62E-03	0.99 ± 0.01	0.62 ± 0.	9.67E-09	1.0 ± 0.00	1.0 ± 0.00
0.85	0.76 ± 0.08	0.65 ± 0.1	1.02E-04	0.98 ± 0.03	0.63 ± 0.01	4.57E-06	1.0 ± 0.00	1.0 ± 0.00
0.95	0.87 ± 0.04	0.67 ± 0.09	1.19E-03	1.00 ± 0.00	0.63 ± 0.01	1.39E-08	1.0 ± 0.00	1.0 ± 0.00
1.00	0.92 ± 0.04	0.71 ± 0.06	2.46E-04	1.00 ± 0.00	0.64 ± 0.01	4.02E-08	1.0 ± 0.00	1.0 ± 0.00

Note. *P*-value of paired *t*-test between KG-SCCA and PMA results are also shown.

in the PMA package (http://cran.r-project.org/web/packages/PMA/) (Witten et al., 2009). For PMA experiments, the SCCA parameters were automatically tuned using a permutation scheme provided in PMA. Below we report our empirical results using both synthetic data and real imaging genetics data.

4.1 Results on simulation data

Because it was not straightforward to manually construct a dataset with a network structure, we simulated group structures for both datasets and then converted them into network structures for one dataset by connecting all the pairs within each group. Synthetic data ($\mathbf{n} = 200$, $\mathbf{p} = 200$, $\mathbf{q} = 150$) with diagonal block structure was generated with the following procedure: (i) Random positive definite covariance matrix **M** with non-overlapping group structure was created, where correlations range from 0.6 to 1 within group and are set to 0 between groups. (ii) Dataset **X** with covariance structure **M** was calculated through Cholesky decomposition. (iii) Repeat Steps 1 and 2 to generate another dataset **Y**. (iv) With assigned canonical loadings of **X**, we calculated the first component **Xu**. (v) Given a desired correlation between components, we calculated the second component **Yv**. (vi) For simplicity, in this article, only one group in **Y** was assigned to have signals. Therefore, based on predefined canonical loadings of **Y** and component **Yv**, final obtained group signals, added with some white noises (Signal to Noise Ratio (SNR) = 0.5), will replace the data in original dataset **Y**. By repeating this procedure we generated seven datasets with correlation levels from 0.6 to 1. The canonical loadings and group structure remained the same across all the datasets.

KG-SCCA and PMA have been both tested on all seven datasets. All the regularization parameters were optimally tuned using a grid search from 10^{-2} to 10^{2} through nested 5-fold cross-validation, as mentioned before. The true and estimated canonical loadings for both **X** and **Y** were shown in Figure 3. Owing to the difference in normalization and optimization procedure, the weights yielded by KG-SCCA and PMA showed different scales. Yet, the overall profile of the estimated **u** and **v** values from KG-SCCA kept consistent with the ground truth across the entire range of tested correlation strengths (from 0.6 to 1.0), whereas PMA was only capable of identifying an incomplete portion of all the signals. Furthermore, we also examined the correlation in the test set computed using the learned models from the training data for both methods. The left part of Table 2 demonstrated that KG-SCCA outperformed PMA consistently

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Table 3. Five-fold cross validation results on real data: the models learned from the training data were used to estimate the correlation coefficients between canonical components for both training and testing sets

Method		Train						Test					
		f1	f2	f3	f4	f5	Mean	f1	f2	f3	f4	f5	Mean
KG-SCCA	exp1	0.471	0.448	0.475	0.451	0.46	0.461	0.431	0.515	0.401	0.417	0.459	0.445
	exp2	0.476	0.453	0.454	0.476	0.461	0.464	0.402	0.505	0.503	0.401	0.458	0.454
	exp3	0.476	0.474	0.474	0.468	0.402	0.459	0.408	0.393	0.413	0.435	0.565	0.443
	exp4	0.468	0.466	0.459	0.46	0.466	0.464	0.441	0.409	0.47	0.476	0.445	0.448
	exp5	0.49	0.502	0.434	0.449	0.447	0.464	0.35	0.297	0.584	0.527	0.528	0.457
PMA	exp1	0.439	0.418	0.438	0.438	0.426	0.432	0.368	0.45	0.398	0.379	0.439	0.407
	exp2	0.444	0.416	0.425	0.436	0.432	0.431	0.354	0.463	0.449	0.399	0.416	0.416
	exp3	0.442	0.445	0.439	0.427	0.398	0.43	0.382	0.341	0.382	0.432	0.544	0.416
	exp4	0.434	0.44	0.425	0.427	0.431	0.432	0.414	0.363	0.445	0.438	0.415	0.415
	exp5	0.459	0.462	0.406	0.416	0.411	0.431	0.288	0.287	0.517	0.486	0.501	0.416
						P-value	3.08E-6					*P*-value	8.07E-5

Note. *P*-values of paired *t*-tests were obtained for comparing KG-SCCA and PMA results.

and significantly, and it could accurately reveal the embedded true correlation even in the test data. The right part of Table 2 demonstrated the sensitivity and specificity performance using area under ROC (AUC), where KG-SCCA also significantly outperformed PMA no matter whether the correlation was weak or strong in **u**. Because **v** is relatively simple structured, both KG-SCCA and PMA can restore the signals without any loss. From the above results, it is also observed that KG-SCCA could identify the correlations and signal locations not only more accurately but also more stably.

4.2 Results on real imaging genetic data

Both KG-SCCA and PMA have been performed on real amyloid imaging and *APOE* genetics data. Similar to previous analysis, 5-fold nested cross-validation was applied to optimally tune the parameters. Five experiments were performed with five different partitions to eliminate the bias. For each single experiment, the same partition was used for both KG-SCCA and PMA. Table 3 shows both the training and test performances of KG-SCCA and PMA in all five folds of five experiments. Both methods demonstrated stable results across five trials. KG-SCCA was observed to outperform the PMA in every single experiment on both training and test performance. Paired *t*-test was performed to compare the performance across five experiments, and KG-SCCA outperformed PMA significantly in both training ($P = 3.08\text{E-}6$) and test cases ($P = 8.07\text{E-}5$). We also tested two simplified KG-SCCA models: one with only the penalty term for the LD structure and the other with only the penalty term for the network structure. Interestingly, both performed similarly to the original KG-SCCA, and significantly outperformed PMA.

Figure 4 demonstrates the canonical loadings trained from 5-fold cross-validation in one experiment, suggesting relevant genetic (top panel) and imaging (bottom panel) markers. Although LD block constraints were imposed on relevant SNP markers, L_1 penalty managed to exclude irrelevant signals. Only *APOE* e4 SNP (rs429358) was identified to be associated with amyloid accumulations in the brain. PMA also achieved a similar pattern as KG-SCCA, but including a few additional SNPs from multiple LD blocks. The bottom panel of Figure 4 shows the canonical loading for the imaging data. Both methods identified similar imaging patterns, which are in accordance with prior findings (Ramanan *et al.*, 2014). Figure 5 shows a brain map of canonical loadings generated by KG-SCCA.

5 CONCLUSIONS

We have performed a brain imaging genetics study to explore the relationship between brain-wide amyloid accumulation and genetic variations in the *APOE* gene. Because most existing SCCA algorithms are designed using the soft thresholding technique, which assumes independence among data features, direct application of these methods into brain imaging genetics study cannot yield optimal results owing to the correlated imaging and genetic features. We have proposed a novel KG-SCCA algorithm, which not only removes the above independence assumption, but also can model both the group-like and network-like prior knowledge in the data to produce improved learning results. A comparative study has been performed between KG-SCCA and PMA (a widely used SCCA implementation) on both synthetic and real data. The promising empirical results demonstrated that KG-SCCA significantly outperformed PMA in both cases. Furthermore, KG-SCCA could accurately recover the true signals from the synthetic data, as well as yield improved canonical correlation performances and biologically meaningful findings from real data. This study is an initial attempt to remove the feature independence assumption many existing SCCA methods have. The empirical studies designed here are targeted to identify relatively clean and simple multi-SNP–multi-QT correlations. Given only 58 SNPs analyzed here, this work is not a demonstration of a genome-wide analysis. Comparison with other complex SCCA models, building scalable KG-SCCA models, and applications to more complex imaging genetic tasks warrant further investigation.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

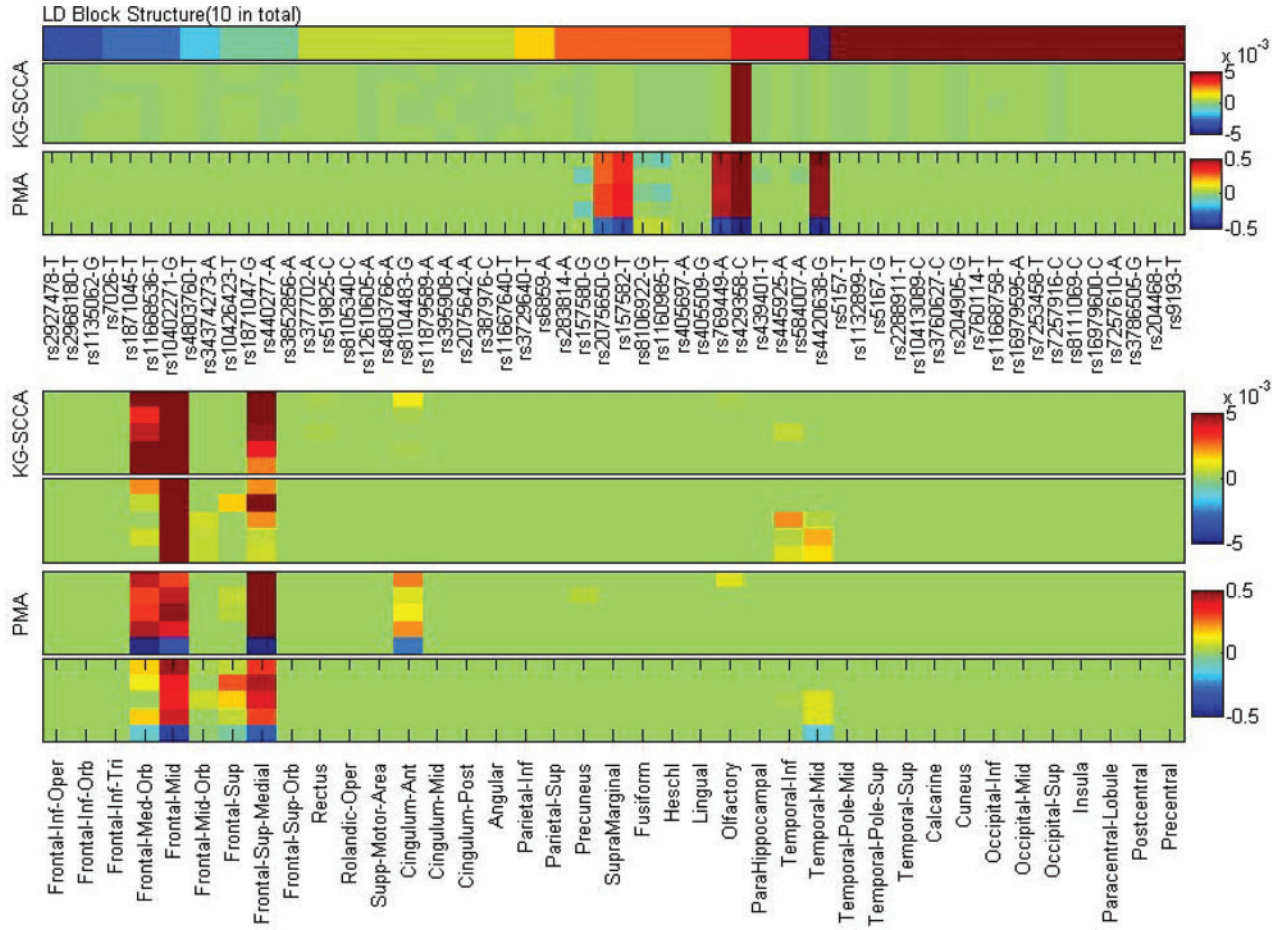


Fig. 4. Five-fold trained weights of **u** (top panel) and **v** (bottom panel). KG-SCCA results and PMA results are shown for each panel. For each of KG-SCCA and PMA imaging results (i.e. the bottom panel), the top and bottom rows correspond to left and right hemispheres, respectively

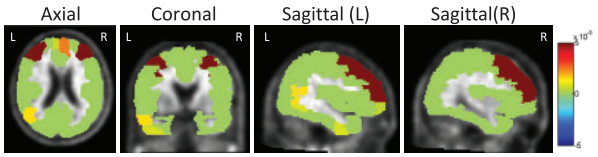


Fig. 5. Mapping canonical loading generated by KG-SCCA onto the brain

ACKNOWLEDGEMENTS

Detailed ADNI Acknowledgements information is available in http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Manuscript_Citations.pdf. 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.

Funding: This work was supported by the National Institutes of Health [R01 LM011360 to L.S. and A.S., U01 AG024904 to M.W. and A.S., RC2 AG036535 to M.W. and A.S., R01 AG19771 to A.S., P30 AG10133 to A.S.] and the National Science Foundation [IIS-1117335 to L.S.] at IU; by the National Science Foundation [IIS-1117965 to H.H., IIS-1302675 to H.H., IIS-1344152 to H.H., DBI-1356628 to H.H.] at UTA; and by the National Institutes of Health [R01 LM011360 to J.M., R01 LM009012 to J.M., R01 LM010098 to J.M.] at Dartmouth.

Conflict of interest: none declared.

REFERENCES

Barrett,J.C. (2009) Haploview: visualization and analysis of SNP genotype data. *Cold Spring Harb. Protoc.*, **2009**, pdb ip71.

Chen,J. *et al.* (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.

Chen,X. and Liu,H. (2012) An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. *Stat. Biosci.*, **4**, 3–26.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Chi,E. *et al.* (2013) Imaging genetics via sparse canonical correlation analysis. In: *Biomedical Imaging (ISBI), 2013 IEEE 10th Int Sym on*. San Francisco, USA, pp. 740–743.

Hibar,D.P. *et al.* (2011) Multilocus genetic analysis of brain images. *Front. Genet.*, **2**, 73.

Jagust,W.J. *et al.* (2010) The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. *Alzheimers Dement*, **6**, 221–229.

Lin,D. *et al.* (2014) Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Med. Image Anal.*, **18**, 891–902.

Ramanan,V.K. *et al.* (2014) APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. *Mol. Psychiatry*, **19**, 351–357.

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.

Swaminathan,S. *et al.* (2012) Amyloid pathway-based candidate gene analysis of [(11)C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain Imaging Behav.*, **6**, 1–15.

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

Vounou,M. *et al.* (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *NeuroImage*, **53**, 1147–1159.

Wan,J. *et al.* (2011) Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. *Med. Image Comput. Comput. Assist. Interv.*, **14** (Pt 2), 376–383.

Wang,H. *et al.* (2012) Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*, **28**, 229–237.

Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, **8**, Article28.

Witten,D.M. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yan,J. *et al.* (2013) Network-guided sparse learning for predicting cognitive outcomes from MRI measures. In: *Multimodal Brain Image Analysis (MBIA), Nagoya, Japan*. LNCS 8159. Vol. 8159, Springer International Publishing, Switzerland, pp. 150–158.

Zeng,H. *et al.* (2012) Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, **149**, 483–496.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014