Published in final edited form as:
Med Image Anal. 2014 August ; 18(6): 891–902. doi:10.1016/j.media.2013.10.010.

Correspondence between fMRI and SNP data by group sparse canonical correlation analysis

Dongdong Lin[a,b], **Vince D. Calhoun**[c], and **Yu-Ping Wang**[a,b,d,*]

[a] Biomedical Engineering Department, Tulane University, New Orleans, LA 70118, USA

[b] Center of Genomics and Bioinformatics, Tulane University, New Orleans, LA 70118, USA

[c] The Mind Research Network, Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA

[d] Center for Systems Biomedicine, Shanghai University for Science and Technology, Shanghai, China

Abstract

Both genetic variants and brain region abnormalities are recognized as important factors for complex diseases (*e.g.*, schizophrenia). In this paper, we investigated the correspondence between single nucleotide polymorphism (SNP) and brain activity measured by functional magnetic resonance imaging (fMRI) to understand how genetic variation influences the brain activity. A group sparse canonical correlation analysis method (group sparse CCA) was developed to explore the correlation between these two datasets which are high dimensional-the number of SNPs/voxels is far greater than the number of samples. Different from the existing sparse CCA methods (sCCA), our approach can exploit structural information in the correlation analysis by introducing group constraints. A simulation study demonstrates that it outperforms the existing sCCA. We applied this method to the real data analysis and identified two pairs of significant canonical variates with average correlations of 0.4527 and 0.4292 respectively, which were used to identify genes and voxels associated with schizophrenia. The selected genes are mostly from 5 schizophrenia (SZ)-related signalling pathways. The brain mappings of the selected voxles also indicate the abnormal brain regions susceptible to schizophrenia. A gene and brain region of interest (ROI) correlation analysis was further performed to confirm the significant correlations between genes and ROIs.

Keywords

Group sparse CCA; SNP; fMRI; Feature selection; Imaging genetics

* Corresponding author at: Biomedical Engineering Department, Tulane University, New Orleans, LA, USA. wyp@tulane.edu (Y.-P. Wang)..
dlin5@tulane.edu (D. Lin), vcalhoun@unm.edu (V.D. Calhoun)

1. Introduction

Schizophrenia is a complex disease and considered to be caused by the interplay of a number of genetic factors (*e.g.*, change of gene regulation, and alteration of mRNA and SNP) and environmental effects. Genetic factors play an important role in causing schizophrenia disease. People born from a family with a history of schizophrenia have higher risks of schizophrenia than those without a family schizophrenia history. In recent years, many studies have focused on exploring critical genes associated with the schizophrenia. Many potential genetic variants have been reported as possible risk factors such as the G72/G30 gene locus on chromosome 13q (Badner and Gershon, 2002; Abecasis et al., 2004) Gene DISC1 variation (Callicott et al., 2005; Porteous et al., 2006) and copy number variations on gene GRIK3, EFNA5, AKAP5 and CACNG2 (Wilson et al., 2006; Sutrala et al., 2007). In addition to genetic studies, fMRI has also been widely used for the study of schizophrenia because of its capability to identify functional abnormalities within brain regions of schizophrenic patients (Jansma et al., 2004; Li et al., 2007; Meda et al., 2008; Szycik et al., 2009).

Genetic variants and brain region abnormalities are both important markers for the study of schizophrenia. Combining both data can not only contribute to a better understanding of biological mechanisms on brain structure and function but also have the potential to improve the diagnosis and treatments of complex diseases. However, current imaging genetics studies either take brain imaging measurements as endophenotypes to study the associated genetic variants or investigate the effects of a small set of candidate genetic variants on the whole brain measurements (Hamid et al., 2009; Le Cao et al., 2009; Wiley, 2011). It is still challenging to explore the relationship between a large amount of genetic variants and a large number of brain imaging measurements. Therefore, correlative analysis approaches for large-scale multi-modal data analysis are highly demanded.

In this work, we aim to study the effects of multiple SNPs or genes on functional brain activity in schizophrenia. An effective multivariate statistical method is needed. Canonical Correlation Analysis (CCA (Hotelling, 1936)) or Partial Least Squares regression (PLSR (Le Cao et al., 2008)) have been proposed to analyze multi-modal datasets. The CCA aims to maximize the correlation between the linear combinations of variables from two data sets, *e.g.*, a linear combination of SNPs and a linear combination of voxels. However, the method will have the over-fitting issues in analyzing high dimensional data such as SNP and brain imaging data as shown in Fig. 1. Thousands of SNPs with linkage disequilibrium (LD) are detected to reflect the genetic variant at different locus. The number of voxels included in the whole brain fMRI image is also very large (*e.g.*, $53 \times 63 \times 46$). Traditional CCA will perform poorly in such a case due to the multi-collinearity (linear dependence) problem, and thus having computational difficulty (Park-homenko et al., 2009). To address above issue, sparse CCA (sCCA) methods, mostly using the $l - 1$ norm (*CCA-l1*) or the combination of $l - 1$ and $l - 2$ norm (*CCA-elastic net*) penalties, have been developed by introducing the sparse penalties into the traditional CCA model (Waaijenborg et al., 2008; Le Cao et al., 2009; Parkhomenko et al., 2009; Witten et al., 2009; Witten and Tibshirani, 2009; Boutte and Liu, 2010). Despite the success, they didn't account for group structures within the data in the analysis (*e.g.*, multiple SNPs within the same gene, a group of voxels within the same

region, a group of voxels within the same ROI, etc.), which often exist or are implied by the biological mechanism. For example, SNPs within the same gene have similar functions and act together at the gene or pathway level to affect the brain activity. These SNP effects can be added up to a larger difference (Tyekucheva et al., 2011). Several previous works have shown the benefit of accounting for the group effect of features in the sCCA models (Chen and Liu, 2012; Chen et al., 2013; Lin et al., 2013). However, to our knowledge, little work has been reported to incorporate the group effect into the sCCA model for fMRI and SNP data integration. Motivated by this fact, in this paper, we developed a group sparse CCA model based integration method by imposing the sparse group lasso penalty on the CCA model for the integrative analysis of SNP and fMRI data; please refer to Fig. 1 for illustration. This method has the following advantages: (1) A group of features (voxels/SNPs) will be inspected during the correlation analysis, which can study the joint effects of multiple SNPs on the regions of voxels; (2) feature selection will be performed at both group and single feature level. Irrelevant groups of features as well as single feature within each group can be removed. Our group sparse CCA method can both exploit group information in the correlation analysis while filter out noisy features within the group simultaneously.

The group sparse CCA can estimate the correlation between canonical variates, corresponding to a set of significant SNPs or brain imaging voxels. Based on the estimates, we provided a gene-ROI correlation analysis to further confirm the significance of the correlations between genes and brain functions in ROIs.

The rest of the paper is organized as follows. The proposed group sparse CCA model and algorithm are introduced in the section of theory. The group sparse CCA based integration method for SNP and fMRI data is described in the section of method. The validation and comparison of our model with other sCCA models on both simulated and real data analysis are presented in the section of results. The pathway analysis results and limitations of the proposed method are discussed finally.

2. Theory

In this section, we first introduces CCA model, based on which the group sparse model is presented. Then we propose a numerical algorithm based on block coordinate descent to solve the model. Finally, we show that the general model we propose can include several existing sCCA models and hence the numerical algorithm can also be applied for their efficient solutions.

2.1. Group sparse CCA

We consider two sets of data $\boldsymbol{X}$ and $\boldsymbol{Y}$ with n samples, where $\boldsymbol{X}$ has p variables and $\boldsymbol{Y}$ has q variables. Assuming $p, q \gg n$ and the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been standardized such that the mean of each column is zero and the 2-norm is one. The variance matrices and covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted by $\sum_{XX}, \sum_{YY}, \sum_{XY}$ (or $\sum_{YX}$) respectively. The CCA model aims to find two loading vectors or projections $\boldsymbol{\alpha}, \boldsymbol{\beta}$, to maximize the correlation between the linear combinations of variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{\alpha}^t\boldsymbol{X}$ and $\boldsymbol{\beta}^t\boldsymbol{Y}$ as shown in the following equation:

$$\max_{\boldsymbol{\alpha},\beta} \boldsymbol{\alpha}^t \sum_{\boldsymbol{XY}} \boldsymbol{\beta} \quad \mathbf{s.t.} \boldsymbol{\alpha}^{\mathbf{t}} \sum_{\boldsymbol{XX}} \boldsymbol{\alpha} = \mathbf{1}, \qquad \boldsymbol{\beta}^{\mathbf{t}} \sum_{\boldsymbol{YY}} \boldsymbol{\beta} = \mathbf{1} \quad (1)$$

There are over-fitting issues in Eq. (1) due to the small samples but high dimensional variables (i.e., $p, q \gg n$), which also might result in the ill-condition of the matrices $\sum_{XX}$ and $\sum_{YY}$. So, sparse penalties such as the l – 1 norm, elastic net, have been imposed on the loading vectors in Eq. (2) in CCA analysis.

$$\min_{\boldsymbol{\alpha},\beta} -\boldsymbol{\alpha}^t \sum_{\boldsymbol{XY}} \boldsymbol{\beta} + \lambda_1 \boldsymbol{\Psi}(\boldsymbol{\alpha}) + \lambda_2 \boldsymbol{\Phi}(\boldsymbol{\beta}) \quad \mathbf{s.t.} \boldsymbol{\alpha}^{\mathbf{t}} \sum_{\boldsymbol{XX}} \boldsymbol{\alpha} \leq \mathbf{1}, \quad \boldsymbol{\beta}^{\mathbf{t}} \sum_{\boldsymbol{YY}} \boldsymbol{\beta} \leq \mathbf{1} \quad (2)$$

where $\boldsymbol{\Psi}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ denote the penalized function on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, which are often taken to be l – 1 norm, $\boldsymbol{\Psi}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$, $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. These sparse penalties can result in a large number of features/variables to zero in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In addition, the constrained conditions in Eq. (2) are relaxed compared to Eq. (1) so that we can first find the alternative solution in a closed region and then obtain the solution of the optimization problem satisfying $\boldsymbol{\alpha}^t \sum_{XX} \boldsymbol{\alpha} = \mathbf{1}$, $\boldsymbol{\beta}^t \sum_{YY} \boldsymbol{\beta} = \mathbf{1}$ (see optimization algorithms in Section 2.2). Similar to conventional CCA, Sparse CCA can also perform multiple pairs of canonical variates extraction iteratively as the following procedure: maximizing the correlation between two data sets by extracting the first pair of sparse loading vectors as well as pair of canonical variates; then removing the effects of the first pair of canonical variates and finding the second sparse loading vectors that maximizes the correlation but is irrelevant to the first pair. This process will not stop until the rth projection pair is gained ($r = rank$ ($\boldsymbol{X}^t\boldsymbol{Y}$)).

Group information of variables is expected to be incorporated in the Eq. (2) by changing the penalized function since in many applications, a set of variables often form a group (*e.g.*, SNPs spanning a gene, genes in a pathway). Therefore, we propose a group sparse CCA model by introducing the sparse group lasso penalty into Eq. (2). For simplicity, we will only consider non-overlapping groups in this paper and assume that variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are partitioned into L and H disjoint groups respectively. The following model, namely group sparse CCA (or CCA-sparse group), is proposed to consider group structures existed in the data:

$$\min_{\boldsymbol{\alpha},\beta} -\boldsymbol{\alpha}^t \sum_{\boldsymbol{XY}} \boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\alpha}\|_{\mathbf{G}} + \tau_1 \|\boldsymbol{\alpha}\|_{\mathbf{1}} + \lambda_2 \|\boldsymbol{\beta}\|_{\mathbf{G}} + \tau_2 \|\boldsymbol{\beta}\|_{\mathbf{1}} \quad \mathbf{s.t.} \boldsymbol{\alpha}^{\mathbf{t}} \sum_{\boldsymbol{XX}} \alpha \leq \mathbf{1}, \quad \boldsymbol{\beta}^{\mathbf{t}} \sum_{\boldsymbol{YY}} \boldsymbol{\beta} \leq \mathbf{1} \quad (3)$$

where $\|\boldsymbol{\alpha}\|_G = \sum_{l=1}^{L} \omega_l \|\boldsymbol{\alpha}\|_2$, $\|\boldsymbol{\beta}\|_G = \sum_{h=1}^{H} \mu_h \|\boldsymbol{\beta}_h\|_2$ are the group penalties to account for joint effects of features within the same group. This model is more realistic in many cases, *e.g.*, multiple SNPs rather than individual SNPs from the same gene work together as a group to be associated with a disease. The group penalty uses the nondifferentiability of $\|\boldsymbol{\alpha}_l\|_2$ (or $\|\boldsymbol{\beta}_h\|_2$) at $\boldsymbol{\alpha}_l = 0$ ($\boldsymbol{\beta}_h = 0$) to set the coefficients of the group to be 0; then the entire group of features will be removed to achieve the group sparsity. While we consider group effects, we can still keep the selection of individual variable/feature. So the l – 1 norm penalties on the individual features (i.e., $\|\boldsymbol{\alpha}\|_1$ and $\|\boldsymbol{\beta}\|_1$) are imposed. λ_1 and λ_2 are the

tuning parameters to control the group sparsity while τ_1 and τ_2 are used to control individual feature sparsity. ω_l and μ_h are the weights to adjust for the group size differences. We set them to be $s_i^{1/2}$, where s_i is the ith group size.

In a particular case, since sparse group lasso penalty is a combination of $l - 1$ norm and group lasso penalty, Eq. (3) can be reduced to the CCA-group model with only group lasso penalty ($\tau_1 = \tau_2 = 0$) and CCA-$l1$ model without group lasso penalty ($\lambda_1 = \lambda_2 = 0$). As discussed above, CCA-group model can tend to select features group by group and keep all features within a selected group while CCA-$l1$ will ignore group effect among features.

2.2. Iterative optimization algorithm

To solve Eq. (3), we rewrite Eq. (3) based on the singular vector decomposition of matrix $\boldsymbol{K}$ and impose sparse penalties on vectors $\boldsymbol{u}, \boldsymbol{v}$

$$\min_{\boldsymbol{u},\boldsymbol{v}} \|\boldsymbol{K} - d\boldsymbol{u}\boldsymbol{v}^t\|_F^2 + \lambda_1\|\boldsymbol{u}\|_G + \tau_1\|\boldsymbol{u}\|_1 + \lambda_2\|\boldsymbol{v}\|_G + \tau_2\|\boldsymbol{v}\|_1 \quad s.t. \quad \|\boldsymbol{u}\|_2^2 \le 1, \quad \|\boldsymbol{v}\|_2^2 \le 1 \tag{4}$$

where $\boldsymbol{K} = \sum_{\boldsymbol{XX}}^{-1/2} \sum_{\boldsymbol{XY}} \sum_{\boldsymbol{YY}}^{-1/2} = \sum_i d_i \boldsymbol{u}_i \boldsymbol{v}_i$, and d_i is the positive square root of the ith eigenvalue of $\boldsymbol{K}^t\boldsymbol{K}$. $\boldsymbol{u}_i$ $\boldsymbol{v}_i$ are the ith eigenvectors of $\boldsymbol{K}^t\boldsymbol{K}$ corresponding to d_i. $\|u\|_2^2 = 1, \|\boldsymbol{v}\|_2^2 = 1$ should be satisfied when the solution of the optimization problem is obtained. The loading vectors can then be derived by

$$\boldsymbol{\alpha}_i = \sum_{\boldsymbol{XX}}^{-1/2} \boldsymbol{u}_i, \quad \text{and} \quad \boldsymbol{\beta}_i = \sum_{\boldsymbol{YY}}^{-1/2} \boldsymbol{v}_i \tag{5}$$

The matrices $\sum_{\boldsymbol{XX}}$ and $\sum_{\boldsymbol{YY}}$ in Eq. (5) might be ill-conditioned because of the high dimensionality of data. We adopt Witten and Tibshirani's (2009) method by replacing the covariance matrices with identity matrices $\boldsymbol{I}$ and hence penalizing the vectors $\boldsymbol{u}, \boldsymbol{v}$ instead of the loading vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Since the problem in Eq. (4) is still not a convex optimization with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$, for simplicity, we decouple the problem into two simple biconvex optimizations by fixing $\boldsymbol{u}$ and $\boldsymbol{v}$ alternatively. An iterative algorithm is then derived to solve the problem, as shown in Table 1.

Taking the solution of (a) for example, one can find the solution with the following Lagrange form formulation:

$$\min_{\boldsymbol{u}} -tr\left(\boldsymbol{K}\boldsymbol{v}\boldsymbol{u}^t\right) + \lambda_1 \sum_{l=1}^{L} \omega_l \|\boldsymbol{u}_l\|_2 + \tau_1|\boldsymbol{u}|_1 + \Delta\left(\|\boldsymbol{u}\|_2^2 - 1\right) \tag{6}$$

where Δ is the parameter to make $\|\boldsymbol{u}\|_2^2 = \sum_{l=1}^{L} \|\boldsymbol{u}_l\|_2^2 = 1$. This is converted into a simple optimization problem with separable objection function and sparse group lasso penalty. A

block coordinate decent algorithm has been developed to solve this problem efficiently as shown in Table 2.

The coordinate decent algorithm has been shown to be effective in solving generalized linear regression models (Wu et al., 2009; Friedman et al., 2010a, 2010b), especially for underdetermined system. It estimates parameter one by one by fixing the other parameters unchanged. Similarly, instead of estimating parameter individually, block coordinate decent algorithm will estimate a block of parameters each time while fixing the other blocks of parameters. In data set $\boldsymbol{X}$, each group $k = 1,2,\ldots,L$, will be inspected one by one. If a group is selected, we will then select each feature in the group by the coordinate decent with the soft-thresholding. Since the optimization is convex, the optimal solution of Eq. (6) is determined by a sub-gradient equation and will converge to a global minimum.

The CCA-group model is a special case of group sparse CCA when $\tau_1 = 0$; the soft-threshold operator in Step (2) (Table 2) need to be changed to the simpler form:

$$\left\| S\left((\boldsymbol{KV})^{(k)}, \tau_1 \right) \right\|_2 = \left\| (\boldsymbol{KV})^{(k)} \right\|_2 \quad (7)$$

The general sparse CCA formula Eq. (3) that we propose can include a variety of sparse CCA models used before (Le Cao et al., 2009; Witten and Tibshirani, 2009), e.g., CCA-*l*1 (Eq. (8)) and CCA-elastic net (Eq. (9)) models. The coordinate decent algorithm can also be applied to solve both models.

$$\Psi(\boldsymbol{u}) = \|\boldsymbol{u}\|_1, \quad \Phi(\boldsymbol{v}) = \|\boldsymbol{v}\|_1 \quad (8)$$

$$\Psi(\boldsymbol{u}) = (1-\delta_1)\|\boldsymbol{u}\|_2^2 + \delta_1\|\boldsymbol{u}\|_1, \Phi(\boldsymbol{v}) = (1-\delta_2)\|\boldsymbol{v}\|_2^2 + \delta_2\|\boldsymbol{v}\|_1 \quad (9)$$

where δ_1, δ_2 are the parameters, controlling the trade-off between $l-2$ norm and $l-1$ norm penalties.

3. Method

We applied group sparse CCA to investigate the association of functional brain regions with genetic variations as shown in Fig. 1. Components extracted from fMRI represent brain regions expressing the functional difference in different subjects. Components from SNP data are linear combinations of SNPs from different genes that may have associations with the disease. After preprocessing, the collected SNPs and ROI-based voxels are both still high dimensional with a large number of features compared to the number of samples. We then used the group sparse CCA to estimate two group loading vectors $\boldsymbol{u}$ and v, from which a pair of canonical variates is obtained. The loading vectors for each component reflect the effect size of the features on the correlation. Then $\boldsymbol{u}$ and $\boldsymbol{v}$ were also used to perform gene-ROI correlation analysis to identify the significantly correlated genes and ROIs.

3.1. Group sparse CCA for fMRI and SNP data analysis

We represented fMRI data collected from participants as a set of spatial voxels. These voxels were divided into 116 ROIs based on the aal (automated anatomical labelling) template (Tzourio-Mazoyer et al., 2002). These ROIs were assumed to be spatially independent but the voxels within each ROI may correlate with each other. These voxels were grouped by ROIs so that we can perform the whole brain analysis (Ng and Abugharbieh, 2011). For SNP data, we extracted those SNPs from preselected 74 reported SZ-risk genes. These SNPs were grouped at gene level (Liu et al., 2013). Hence the $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices can be constructed as follows:

$$\begin{aligned} \boldsymbol{X} &= \left[\boldsymbol{G}_{\boldsymbol{X}}^{1}, \boldsymbol{G}_{\boldsymbol{X}}^{2}, \ldots, \boldsymbol{G}_{\boldsymbol{X}}^{L}\right], \quad \boldsymbol{G}_{\boldsymbol{X}}^{k} \in R^{n \times l_k}, \quad k=1,2,\ldots,L \\ \boldsymbol{Y} &= \left[\boldsymbol{G}_{\boldsymbol{Y}}^{1}, \boldsymbol{G}_{\boldsymbol{Y}}^{2}, \ldots, \boldsymbol{G}_{\boldsymbol{Y}}^{H}\right], \quad \boldsymbol{G}_{\boldsymbol{Y}}^{t} \in R^{n \times h_t}, \quad t=1,2,\ldots,H. \end{aligned}$$

where $L = 74$ indicates the number of genes in SNP data; $H = 116$ is the number of ROIs used in fMRI data, and n is the number of samples. l_k and h_t are the number of SNPs and voxels contained in the kth gene and tth ROI respectively. Under this data structure, two sparse loading vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (or $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) were also grouped in the same way. This way the original data were projected into a low dimensional space.

Four tuning parameters (λ_1; τ_1; λ_2; τ_2) were used in the model to control the group sparsity (λ_1; λ_2) and individual feature sparsity (τ_1, τ_2). A cross-validation can be used to select the optimal parameters but is time-consuming. Therefore, we divided the cross-validation into two steps: (1) using the CCA-group method to select optimal λ_1; λ_2 by considering only the group penalty; (2) selecting the optimal τ_1, τ_2 based on the selected λ_1; λ_2: In summary, the procedure of performing group sparse CCA on the data analysis is as follows:

1. Decompose matrix $\boldsymbol{K}$ in Eq. (5) using SVD to initialize the loading vectors and normalize the vectors with the $l - 2$ norm. Use the two-step cross validation to obtain the optimal tuning parameters.
2. Perform the sub-optimization analysis with respect to each loading vector ($\boldsymbol{u}$ or $\boldsymbol{v}$) in each modality alternatively by the block coordinate decent algorithm. A stopping criterion is assessed for each modality.
3. Two loading vectors are updated and the stopping criteria for both loading vectors are assessed. If they are satisfied, the solution is obtained, go to Step 4. Otherwise, go back to Step 2.
4. Test the significance of the predict correlation ($corr_j$) by permutation. If the correlation is significant, go to Step 5. Otherwise, there is no significant correlation between the current two data sets.
5. Perform gene-ROI analysis to explore the pair-wise correlation between each gene and ROI.

6. Calculate the remaining matrix and repeat step (1–4) to extract the next pair of loading vectors until r pairs of loading vectors are obtained or the extracted predict correlation ($corr_j$) is not significant.

3.2. Tuning parameters selection

The k-fold cross validation was recommended by Waaijenborg and Zwinderman (2009) and Parkhomenko et al. (2009) for parameter selection. Parkhomenko et al. (2009) adopted a criterion that maximizes the mean absolute canonical correlation value of the testing set as shown in the following equation:

$$\Delta_{corr1}=\frac{1}{k}\sum_{i=1}^{k}|cor\left(\boldsymbol{X}_i\,\hat{\boldsymbol{u}}^{-i},\,\boldsymbol{Y}_i\,\hat{\boldsymbol{v}}^{-i}\right)| \quad (10)$$

Waaijenborg et al. (2008) considered the mean difference between the canonical correlations of the training and testing subsets as in the following equation:

$$\Delta_{corr2}=\frac{1}{k}\sum_{i=1}^{k}|cor\left(\boldsymbol{X}_{-i}\,\hat{\boldsymbol{u}}^{-i},\,\boldsymbol{Y}_{-i}\,\hat{\boldsymbol{v}}^{-i}\right)-cor\left(\boldsymbol{X}_i\,\hat{\boldsymbol{u}}^{-i},\,\boldsymbol{Y}_i\,\hat{\boldsymbol{v}}^{-i}\right)| \quad (11)$$

This criterion determines the number of variables which tend to have the same correlations in both training and testing subsets. It is sensitive to the correlation sign change that if the correlation of testing subset changes sign it would be penalized more than when the sign would not change.

Witten and Tibshirani (2009) proposed another permutation based method for optimizing parameters. The whole data set (including both training and testing data sets) will be used to compute the non-permuted correlation d_0 and then $\boldsymbol{X}$ is permuted with T times to calculate the permuted correlation $d_{i=1,\ldots,T}$. The parameters having the largest p-value as in Eq. (12) will be selected.

$$p_{permute}=\left(d_0-\frac{1}{T}\sum_i d_i\right)/std\left(d_{i=1,\ldots,T}\right) \quad (12)$$

Here, $\hat{\boldsymbol{u}}^{-i}$ and $\boldsymbol{v}^{-\hat{i}}$ are estimated loading vectors from training data set. $\boldsymbol{X}_{-i}$ and $\boldsymbol{Y}_i$ are training data set in which subset j is deleted. $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$ are the testing data set. Based on these three criteria, we performed a simulation study using our proposed group sparse CCA model and other sCCA models to choose the optimal criterion (the details are described in Section 4.1).

3.3. Gene-ROI correlation analysis

After the parameters were determined, we obtained the estimates of vectors $\boldsymbol{u}$, $\boldsymbol{v}$ as well as the corresponding canonical variates. We used these estimates to measure the strength and significance of the correlation between genes and ROIs. The null hypothesis of no correlation between a gene and ROI can be written as $H_0 : \rho_{11} = \rho_{12} = \cdots \rho_{ck} = \rho_{s_i s_j} = 0$ versus the alternative hypothesis $H_A : \exists c; k > 0; \rho_{ck} \neq 0$, where s_i and s_j are the number of SNPs and voxels in Gene_i and ROI_j respectively; ρ_{ck} is the correlation between the cth SNP

and the *k*th voxel. To test the hypothesis, we first calculated the pair-wise correlation between the SNPs from Gene_i and voxels from ROI_j. SNPs and voxels were indicated by non-zero coefficients in $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively Then, to avoid the potential bias due to varying gene or ROI size, we averaged the correlation of each SNP-voxel pair as the test statistics as follows.

$$\rho_{ij}=\frac{1}{S_i S_j}\sum_{c=1}^{s_i}\sum_{k=1}^{s_j}|\rho_{ck}| \quad (13)$$

By comparing our observed statistic ρ_{ij} with the null statistics ρ_{ij}^0 with T times permutations of the samples, we can evaluate the significance of the correlation by

$$p \quad value=\left\{\rho_b^0 \geq \rho_{ij}; b=1,2,\ldots,T\right\}/T \quad (14)$$

This gene-ROI correlation analysis can also be applied to the gene–gene correlation and ROI-ROI correlation studies.

3.4. Statistical evaluations

Although our CCA-group sparse method aims to obtain a higher correlation with fewer features selected, in our real data both number of voxels and SNPs are much larger than that of subjects. In order to test the significance of the correlation obtained by the CCA-sparse group method, we took the permutation-based testing. We first obtained $\boldsymbol{u}$ and $\boldsymbol{v}$ based on the optimal parameters, and then permutated the order of subjects in SNP data randomly while keeping the subject order in fMRI data unchanged. The correlation for each permutation can be calculated by the same $\boldsymbol{u}$ and $\boldsymbol{v}$, which is expected to estimate the null distribution of the correlation. The *p*-value can be estimated by large number of permutations. The lower the *p*-value, the more significant the detected correlation is.

4. Results

4.1. Simulation

To assess the performance of the proposed group sparse CCA method, we first simulated two correlated data sets and then we compared group sparse CCA with the other penalized CCA methods such as the CCA-group and CCA-*l*1 on these simulated data.

Two data sets of SNP data $\boldsymbol{X}$ with p SNPs and fMRI data $\boldsymbol{Y}$ with q voxels were simulated. To correlate the SNPs with the voxels, a latent model similar to (Parkhomenko et al., 2009) was used. We first set a latent variable $\Upsilon = \{\gamma_i | i = 1, \ldots, n\}$ with normal distribution $N\left(0, \sigma_\gamma^2\right)$ to have the similar effect on the associated SNPs and voxels in the two data sets. Then $\boldsymbol{X}$ and $\boldsymbol{Y}$ data set were simulated from a multivariate normal distribution by $\boldsymbol{X}^t\boldsymbol{X}$ and $diag\left(\sum_{XX}\right)$ respectively, where the vectors $x_i \in R^p$, $y_i \in R^q$ are the observations of the *i*th sample in $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_j, \ldots, \alpha_p]$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k, \ldots, \beta_q]$, $\alpha_j \neq 0$, $\beta_k \neq 0$, if x_j, y_k are the correlated variables; otherwise, the variables would be considered as noise with zero means ($\alpha_j = 0$, $\beta_k = 0$). γ_i is the ith observation of Υ. σ_γ^2, σ_e^2, are the variances of Υ and noise

variable, which were used to control the correlation between SNPs and voxels. $\sum_{p\times p}$ and $\sum_{q\times q}$ were the variance–covariance matrices of each data set. They were used to simulate the group effect within each data set. For each data set, we set the correlation between the correlated variables within the same group as from the uniform distribution $U(0.1,0.3)$.

The samples of two data sets were simulated from the multivariate normal distribution. Then SNP data $\boldsymbol{X}$ will be converted into categorical variables with three levels (*e.g.*, –1, 0, 1) by specifying the minor allele frequency (MAF) p from the uniform distribution Unif(0.2,0.4). The probability of an observation being from a particular level was set to be p^2, $p(1-p)$ and $(1-p)^2$ respectively. For each variable, the expected number of observations in each level was $n_1 = p^2n$, $n_2 = p(1-p)n$, $n_3 = (1-p)^2n$. We sorted the observations in each variable and assign value 1 to those n_1 observations with the highest value, –1 to those n_3 observations with lowest value and 0 in between (Simon and Tibshirani, 2012). To simulate the fMRI data $\boldsymbol{Y}$ with a brain structure similar to that of real data, we referred to the standard aal template with 116 brain ROIs containing 54277 voxels. The group structure of voxels is based on the ROIs. Voxels correlated with SNPs were generated by latent variable model.

Irrelevant voxels within each ROI were simulated from $y_t \sim N\left(\mu_0^t, \sum_0^t\right)$ where μ_0^t and $\sum_0^t$ are mean and covariance matrix estimated from the tth ROI in real data.

First, we compared the three parameter optimization criteria by applying them to the simulation study. Two data sets $\boldsymbol{X}$ with 200 SNPs and $\boldsymbol{Y}$ with 200 voxels were simulated. Both data sets were divided into 20 groups with group size 10. 20 SNPs from 4 groups in set $\boldsymbol{X}$ and 30 voxels from 5 groups in set $\boldsymbol{Y}$ were set to be correlated. Standard deviation $\sigma_\gamma = 1$ and $\sigma_e = 0.3$. The sample size was 100 and 50 simulations were replicated. The total true positive (TTP) and total discordance (TD: the total false positive plus the total false negative) were used for the comparison. Fig. 2 shows the box-plot of TTP and TD by using different criteria. It can be seen that, by all the methods, Δ_{corr1} tends to give higher TTP but also introduces higher TD; the use of $p_{permute}$ can cause the decrease of TD rapidly at the cost of low TTP; and Δ_{corr2} is a trade-off between these two criteria by keeping a high TTP and decreasing TD to a low level. Hence the last one is chosen as our parameter selection criterion.

Second, we compared the performance of three methods in the simulation. SNP data $\boldsymbol{X}$ with 400 SNPs are divided into 20 groups evenly. 4 groups were randomly selected and 15 SNPs randomly selected from these 4 groups were generated to be correlated with voxels, as shown in Fig. 3(a). For the sake of simplicity, 5% voxels randomly from the first 4 regions with totally 214 voxels were associated with SNPs (Fig. 3(e)). Fig. 3(b–d, f–h) shows the results of recovered loading vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ by CCA-$l1$, CCA-group and CCA-sparse group methods. A 5-fold cross-validation was used to select the optimal parameters. From Table 3, it can be seen that the CCA-sparse group method can better estimate true $\boldsymbol{u}$ and $\boldsymbol{v}$ than other two methods. For the SNP data, all three methods can identify true SNPs (60) while CCA-$l1$ has more false positives (36, shown in Fig. 3(b)) than those of CCA-group (20) and CCA-sparse group (4). For the fMRI data, CCA-$l1$ also misses out more true voxels (29 out of 214) in Fig. 3(f) when selecting $\boldsymbol{v}$. The CCA-group, however, can better recover all the

groups with true features. Nevertheless, it selects all features from the group, which results in very high false positive (4097, shown in Fig. 3(g)). CCA-sparse group can not only identify the group structure to find more true variables than CCA-*l*1 but also remove those noisy features in the group, leading to the least false positive (113 false voxels versus 912 by CCA-*l*1 and 4097 by CCA-group).

Finally, we evaluated the performance of three models with respect to different noise levels, which affected the values of correlation between two data sets. 200 SNPs from 20 genes were simulated in data set $\boldsymbol{X}$. The group size was 10. 50 voxels from 5 ROIs were randomly selected to be correlated with 20 SNPs in $\boldsymbol{X}$. The sample size was 150. According to the correlation estimate in (Parkhomenko et al., 2009), we simulated different level of correlations between two data sets by changing the standard deviation of the noise variable σ_e from 0.1 to 1 with interval 0.1. From the results in Fig. 4, we can see that when the true correlation increases with the decrease of noise variance, more true variables can be recovered with less total discordance by all three methods. The CCA-group model can recover the most correlated variables but has the highest total discordance. When σ_e is larger than 0.6 (the estimated highest correlation is 0.39), the number of TTP and TD is relatively unchanged. When σ_e decreases, more true variables can be identified by all methods and the TD of CCA-*l*1 and CCA-sparse group decreases rapidly.

4.2. Experimental results

In this study, subject recruitment and data collection were conducted by The Mind Clinical Imaging Consortium (MCIC). Two types of data (SNP and fMRI) were collected from 208 subjects including 92 schizophrenia patients (age: 34 ± 11, 22 females) and 116 healthy controls (age: 32 ± 11, 44 females). All of them were provided written informed consents. Healthy participants were free of any medical, neurological or psychiatric illnesses and had no history of substance abuse. By the clinical interview of patients for DSM IV-TR Disorders (22) or the Comprehensive Assessment of Symptoms and History, patients met criteria for DSM-IV-TR schizophrenia (23). Antipsychotic history was collected as part of the psychiatric assessment.

4.2.1. fMRI data collection and preprocessing—fMRI data were collected during a sensor motor task, a block-design motor response to auditory stimulation. The images were acquired on a Siemens3T Trio Scanner and 1.5 T Sonata with echo-planar imaging (EPI) sequences taking parameters (TR = 2000 ms, TE = 30 ms (3.0 T)/40 ms (1.5 T), field of view = 22 cm, slice thickness = 4 mm, 1 mm skip, 27 slices, acquisition matrix = 64 × 64, flip angle = 90°). Data were pre-processed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm) and were realigned, spatially normalized and resliced to 3 × 3 × 3 mm, smoothed with a 10 × 10 × 10 mm^3 Gaussian kernel, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as repressors. Finally the stimulus-on versus stimulus-off contrast images were extracted with 53 × 63 × 46 voxels and all the voxels with missing measurements were excluded. 116 ROIs were extracted based on the aal brain atlas, which resulted in 41236 voxels left for analysis.

4.2.2. Genotyping and preprocessing—A blood sample was obtained for each participant and DNA was extracted. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmni1-Quad assay covering 1,140,419 SNP loci. Bead Studio was used to make the final genotype calls. PLINK software package (http://pngu.mgh.harvard.edu/~purcell/plink) was used to perform a series of standard quality control procedures, resulting in the final dataset spanning 777,635 SNP loci. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 for 'BB' (no minor allele), 1 for 'AB' (one minor allele) and 2 for 'AA' (two minor alleles)'. SNP with >20% missing data were deleted and missing data were further imputed. SNPs with minor allele frequency <1% were removed. To reflect the influence of genetic variation on brain behavior, SNPs included in top 75 schizophrenia genes listed on the SZGene database (http://www.szgene.org/) were selected for the analysis. This procedure yielded 3082 SNPs, which were annotated with 74 genes. There was no SNP found in the remaining one gene.

4.2.3. Correlation analysis—The fMRI voxels were grouped based on ROIs while SNPs were grouped by genes. Then we applied our CCA-sparse group method to the analysis of the correlation between two data sets. 208 subjects were randomly divided into two subsets: 150 subjects for training and the remaining 58 ones for testing. In training data set, we fit three models: (i) CCA-group lasso, i.e., using group lasso penalty on genes and ROIs, (ii) CCA-$l1$, i.e., only imposing $l - 1$ norm on the effects of all SNPs and voxels, and (iii) CCA-sparse group model, i.e., using both group-level and single feature level regularizations. The optimal parameters were obtained from training data by 5-fold cross validation. The models were estimated as well as the features were selected from the training data using the optimal parameters. Then, these estimated models were applied to test data to predict the correlation between two data sets.

To have a stable feature selection, we performed random sampling from 208 subjects repeatedly for B times, selected the same proportion of subjects for training and test data sets, and fitted three models on each sub-sample. We assumed that those SNPs and voxels selected more frequently are more valuable for exploring the correlation between two data sets. For each sub-sample, $\hat{\boldsymbol{u}}_{\boldsymbol{b}}$, $\hat{\boldsymbol{v}}_{\boldsymbol{b}}$ were estimated, $b = 1,2,\ldots,B$. A measure of feature importance can be computed by frequency of their appearance defined as the selection probability in (15). A set of features with high value will be selected by a cut-off threshold.

$$p_j = \frac{1}{B}\sum_{b=1}^{B} I\left(\hat{t}_j^b \neq 0\right) \quad (15)$$

where $\hat{j}_j^b$ is taken to be $\hat{u}_j^b$ (or $\hat{v}_j^b$), i.e., the loading coefficient of sCCA model corresponding to the jth SNP (or voxel), and $I(\cdot)$ is indicator function.

By the algorithm in Section 2.2, we can derive several pairs of canonical variates by iteratively implementing the CCA decomposition. Here, we only show the first two pairs of canonical variates. $B = 50$ replicates were performed. To compare the performance, the test correlations $cor(\boldsymbol{X}_{test}\hat{\boldsymbol{u}}, \boldsymbol{Y}_{test}\hat{\boldsymbol{v}})$ based on the estimated loading vectors $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ in training data

by three methods were calculated. The numbers of selected features were also used for comparison. Fig. 5(a) shows that CCA-group method chose the largest number of SNPs and voxels compared to the other two methods while CCA-sparse group method selected the least features due to the double sparsity constraints on loading vectors. As shown in Table 4, CCA-sparse group generally performs better than the other two methods by obtaining higher predict correlations with comparable low variance. The low variance of these correlations demonstrates the robustness of the estimation. The correlation variance by CCA-sparse group method is also comparable with that of the other two methods. In addition, the small training and test correlation difference (Δ_{corr} = 0.146, 0.1331, 0.1471 for three methods respectively) indicates the advantage of using criterion Eq. (11) for tuning parameters selection, i.e., the stability of the criterion regardless of the method used.

We further discuss the selected features. Choosing SNPs and voxels by ranking the selection probabilities (p_j in Eq. (14)) with a cut-off threshold 0.3, we summarized those top ranked SNPs and voxels in two pairs of variates by three methods in Fig. 5(b and c). There are a high proportion of overlapped SNPs and voxels selected by these methods. Table 5 lists top ranked SNPs and corresponding genes in two pairs of variates by CCA-sparse group method. The first pair contained 51 SNPs from 16 genes correlated with 756 voxels from 36 ROIs with the average correlation 0.4527, $p < 0.001$. 30 SNPs from genes ERBB4 and MAGI2 and 212 voxels from ROIs 7 and 8 were selected by all methods. The other pair had19 SNPs from 9 genes correlated with 1558 voxels from 29 ROIs, and the average correlation was 0.4292, $p < 0.001$. 3 SNPs from MAGI2 and 251 voxels from ROIs 47, 51 were overlapped by all three methods. Table 6 listed the top ranked brain ROIs in the first and second pair of variates respectively. Fig. 6 shows the brain mapping of genomic correlated ROIs. Several brain regions have been reported to be susceptible to schizophrenia from other neuroimaging studies (Shenton et al., 2001; Kumari et al., 2002; Onitsuka et al., 2004; Torrey, 2007; Bellani et al., 2010; Pinault, 2011). They include superior, middle, and medial frontal gyrus, inferior parietal lobule, superior temporal gyrus, thalamus, parahippocampal gyrus, cingulate gyrus, and lingual gyrus, which provided additional evidences that these disease relevant brain regions may be affected by the correlated genomic variations.

4.2.4. Gene-ROI correlation analysis—Based on the two pairs of canonical variates, we further explored and verified the correlation between genes and ROIs. We found that some regions of brain might be correlated with a set of genes. For each gene-ROI, gene-gene and ROI-ROI correlation, 10000 permutations (as mentioned in Method section) were performed to test the significance. In Fig. 7(a), the absolute values of significant gene-ROI correlation are 0.1862 ± 0.0317 (mean ± SD, $p < 0.005$). As shown in Fig. 8(a), Gene ERBB4, NRG1, MAGI2 and GABRG2 show correlations with some ROIs such as 3, 4, 7, 8 (the index of ROI is defined by the aal template (Tzourio-Mazoyer et al., 2002)) with the correlation (ρ = 0.1822, 0.1483, 0.1335,0.1782, $p < 0.004$) respectively. These ROIs mostly consist of superior, middle, medial front gyrus and pre-central gyrus located at frontal lobe which contains primary motor cortex and has been suspected to have abnormal changes in schizophrenia patients (Honey et al., 2005; Kiehl et al., 2005). ROIs 75, 76, 77, 11 and 13(not shown) are also found to be correlated with gene ERBB4, NRG1, and MAGI2 with

the correlation value of 0.1822, 0.1649 and 0.1541 respectively. These ROIs are mainly located at thalamus (right) which plays a critical role in coordinating the pass of information between brain regions. Many studies show the association between dysfunction of thalamus with schizophrenia (Kiehl and Liddle, 2001; Clinton and Meador-Woodruff, 2004; Sui et al., 2011). These three genes are also correlated with each other, which may have the similar effects on the ROIs. In addition, ROIs 91, 92, 99, 100, 103, 111, 112 are correlated with many genes (*e.g.* GRIN2B, CHL1, ERBB4, NRG1, FOXP2, MAGI1, GABRG2, MAGI2). These ROIs are located at declive of cerebellum and culmen. Their relationship with schizophrenia is not clear yet but several previous publications have reported significant difference of these regions between normal control and schizophrenia patients (Kim et al., 2009) and there are models of schizophrenia which discuss importance of the cerebellum (Andreasen and Pierson, 2008).

In the second pair of canonical variates (Fig. 7(b)), the absolute value of significant gene-ROI correlation is 0.1713 ± 0.0308 (mean ± SD, $p < 0.005$). There are two patterns apparent in the figure. The first pattern corresponds to ROIs 40, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 67, 68, which are located at occipital lobe including lingual gyrus, posterior cingulate, precuneus, parahippocampagyrus and superior, middle and inferior occipital gyrus as shown in Fig. 8(b). Four genes ERBB4, CHL1, GRM3 and MAGI2 were significantly correlated with these ROIs at 0.1776, 0.1356, 0.1762 and 0.1429, $p < 0.005$ respectively. These critical regions have been widely studied and are shown to have potential relationship with schizophrenia (Kiehl et al., 2005; Fransson and Marrelec, 2008; Kim et al., 2009). The second pattern are ROIs 1, 24, 57 from precentral gyrus at parietal lobe corresponding to genes SNPA29, MTHFR, SLC1A and NRG1 with the correlation of 0.2251, 0.1899, 0.1816 and 0.21 respectively.

5. Discussion and conclusion

In this paper, we proposed a novel method to explore the relationship between genomic data and fMRI brain imaging data by considering the group effects of the variables in the data. We introduced the group sparse CCA method and the numerical implementation based on the regularized SVD and block coordinate decent algorithm. The performance of group sparse CCA model was compared with other sCCA models in the simulation study, showing that our group sparse CCA method could better recover the true correlations with lower false positive and total discordance. Then we applied the method to correlation analysis between the SNP data and fMRI imaging data. Two pairs of canonical variates with significant correlations were identified. There are 5 pathways implied by these identified genes, which may involve with the biological processes related to schizophrenia. The SZ-risk genes correlated with brain regions have also been reported to be susceptible to schizophrenia, which further validates the results of our method.

We identified two pairs of canonical variates with significant correlations and further verified those linked components between gene-ROI with significant pair-wise correlation. Those identified genes are from the preselected 74 genes from α database which were reported to be susceptible to SZ. Some of these genes may not directly show group difference in our fMRI data. Therefore, in this section, we focus on those group-

discriminating genetic factors associated with motor response task-related brain function disruption in schizophrenia. We use the difference of minor allele frequency (MAF) between case and control group to evaluate the effects of the SNPs in the gene correlated with ROI component. Significantly higher MAF in case group indicates the positive discriminating effect while lower MAF indicates the negative discriminating effect. Pathway analysis of these important genes is performed through Ingenuity Pathway Analysis (IPA: Ingenuity Systems, http://www.ingenuity.com). We found those selected genes involving in 5 pathways as shown in Fig. 9.

1. Neuregulin signalling pathway: NRG1 and ERBB4. These two genes have significant effects in both pairs of canonical variates. It has been reported that the NRG1-ERBB4 modulates some plausible neurobiological mechanisms, i.e., neuronal plasticity in human brain which may be altered in SZ (Buonanno, 2010). In our results, SNPs ‘rs10090544’, ‘rs16878394’, ‘rs7843384’(NRG1) and’rs16847732’, ‘rs16847769’, ‘rs2008506’ in ERBB4 exhibit higher MAFs in patients;rs2466063(NRG1) and the MAFs in’rs11903508’, ‘rs16846111’, ‘rs16846352’in ERBB4 are lower in patients indicating negative effect.

2. Glutamate receptor signalling: GRIN2B, GRM3 and SLC1A2. These genes are involved in encoding ionotropic glutamate receptors and histamine receptors which may regulate the neurotransmitter transmission associated with SZ (Bishop et al., 2005). Several researches have supported the possibility of GRIN2B conferring the susceptibility to schizophrenia (Ohtsuki et al., 2001; Qin et al., 2005). GRM3 may be involved in the pathophysiology of schizophrenia, and its associated cognitive impairment, especially at prefrontal and hippocampal regions (Egan et al., 2004). In our results, ‘rs2299218’, ‘rs802432’ in GRM3 and ‘rs2284425’ in GRIN2B have lower MAFs in patients showing negative effects while ‘rs3026164’(GRIN2B), ‘rs802425’ (GRM3) and ‘rs11033095’ (SLC1A2) have positive effects on the disease.

3. GABA receptor signalling: GABRB2 and GABRG2. GABA neuronal dysfunction has been found to associate with cognitive impairment of schizophrenia. We identified a SNP ‘rs153303’ (GABRB2) with lower MAF in schizophrenia, showing negative effect while ‘rs211037’ (GABRG2) has higher MAF in patients with a possible opposite influence.

4. Calcium signalling pathway: HTR3A, ERBB4 and PPP3CC. Recentbiochemical research supports that schizophrenia may be produced by alterations in various intracellular molecules as long as these alterations lead to abnormal functioning of some central intracellular regulatory pathways and one of unifying elements of molecular changes in schizophrenia is their association with potential altered Ca^{2+}signalling (Lidow, 2003).‘rs1150219’ (HTR3A) and ‘rs7010861’ (PPP3CC) both show higher MAFs in patients in the results.

5. Tight junction: MAGI1, MAGI2. These genes attend the coding for MAGI proteins which may relate to SZ by influencing the development and communication between nerve cells (Karlsson et al., 2012). In our results, ‘rs9311944’, ’rs6809559’ (MAGI1), and ‘rs10230275’, ’rs10229284’, ’rs1330490’

(MAGI2) present higher MAFs in patients while ‘rs9821646’, ’rs2061937’ (MAGI1), ’rs10230752’, ’rs10953782’(MAGI2) exhibit lower MAFs in patients.

The tuning parameters selection criteria used in this work are also important. The performance of Waaijenborg et al.'s criterion is better than those of other two criteria in terms of feature selection accuracy, which is consistent with the conclusion from (Waaijenborg and Zwinderman, 2009). However, these parameter selection criteria are not directly used for the purpose of feature selection; a better criterion can be developed. In addition, the selection of tuning parameters is via cross-validation, relying on large training data, which is often not available. Alternatively, Bayesian information criterion and Akaike information criterion (Posada and Buckley, 2004) can be used for the optimal parameter selection.

Acknowledgments

This work is partially supported by both NSF and NIH. It is also supported by Shanghai Eastern Scholarship Program.

References

Abecasis GR, Burt RA, et al. Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. Am. J. Hum. Genet. 2004; 74(3):403–417. [PubMed: 14750073]

Andreasen NC, Pierson R. The role of the cerebellum in schizophrenia. Biol. Psychiatr. 2008; 64(2): 81–88.

Badner JA, Gershon ES. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Mol. Psychiatr. 2002; 7(4):405–411.

Bellani M, Ferro A, et al. The potential role of the parietal lobe in schizophrenia. Epidemiol. Psichiatr. Soc. 2010; 19(2):118–119. [PubMed: 20815295]

Bishop JR, Ellingrod VL, et al. Association between the polymorphic GRM3 gene and negative symptom improvement during olanzapine treatment. Schizophr. Res. 2005; 77(2–3):253–260. [PubMed: 15913960]

Boutte, D.; Liu, J. Sparse canonical correlation analysis applied to fMRI and genetic data fusion.; IEEE International Conference on Bioinformatics and Biomedicine; 2010.

Buonanno A. The neuregulin signaling pathway and schizophrenia: from genes to synapses and neural circuits. Brain Res. Bull. 2010; 83(3–4):122–131. [PubMed: 20688137]

Callicott JH, Straub RE, et al. Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia. Proc. Natl. Acad. Sci. USA. 2005; 102(24):8627–8632. [PubMed: 15939883]

Chen X, Liu H. An efficient optimization algorithm for structured sparse CCA, with applications to eQTL Mapping. Stat. Biosci. 2012; 4:3–26.

Chen J, Bushman FD, et al. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics. 2013; 14(2):244–258. [PubMed: 23074263]

Clinton SM, Meador-Woodruff JH. Thalamic dysfunction in schizophrenia: neurochemical, neuropathological, and in vivo imaging abnormalities. Schizophr. Res. 2004; 69(2–3):237–253. [PubMed: 15469196]

Egan MF, Straub RE, et al. Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. Proc. Natl. Acad. Sci. USA. 2004; 101(34):12604–12609. [PubMed: 15310849]

Fransson P, Marrelec G. The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: evidence from a partial correlation network analysis. Neuroimage. 2008; 42(3): 1178–1184. [PubMed: 18598773]

Friedman, J.; Hastie, T., et al. A Note on the Group Lasso and a Sparse Group Lasso. 2010. <http://www.arxiv.org/pdf/1001.0736>

Friedman J, Hastie T, et al. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 2010b; 33(1):1–22. [PubMed: 20808728]

Hamid JS, Hu P, et al. Data integration in genetics and genomics: methods and challenges. Hum. Genom., Proteom. 2009

Honey GD, Pomarol-Clotet E, et al. Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function. Brain. 2005; 128(Pt 11):2597–2611. [PubMed: 16183659]

Hotelling H. Relations between two sets of variates. Biometrika. 1936; 28:321–377.

Jansma JM, Ramsey NF, et al. Working memory capacity in schizophrenia: a parametric fMRI study. Schizophr. Res. 2004; 68(2–3):159–171. [PubMed: 15099600]

Karlsson R, Graae L, et al. MAGI1 copy number variation in bipolar affective disorder and schizophrenia. Biol. Psychiatr. 2012; 71(10):922–930.

Kiehl KA, Liddle PF. An event-related functional magnetic resonance imaging study of an auditory oddball task in schizophrenia. Schizophr. Res. 2001; 48(2–3):159–171. [PubMed: 11295369]

Kiehl KA, Stevens MC, et al. Abnormal hemodynamics in schizophrenia during an auditory oddball task. Biol. Psychiatr. 2005; 57(9):1029–1040.

Kim DI, Mathalon DH, et al. Auditory oddball deficits in schizophrenia: an independent component analysis of the fMRI multisite function BIRN study. Schizophr. Bull. 2009; 35(1):67–81. [PubMed: 19074498]

Kumari V, Gray JA, et al. Procedural learning in schizophrenia: a functional magnetic resonance imaging investigation. Schizophr. Res. 2002; 57(1):97–107. [PubMed: 12165380]

Le Cao KA, Rossouw D, et al. A sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. 2008; 7 Article 35.

Le Cao KA, Martin PGP, et al. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinform. 2009; 10:34.

Li X, Branch CA, et al. FMRI study of language activation in schizophrenia, schizoaffective disorder and in individuals genetically at high risk. Schizophr. Res. 2007; 96(1–3):14–24. [PubMed: 17719745]

Lidow MS. Calcium signaling dysfunction in schizophrenia: a unifying approach. Brain Res. Brain Res. Rev. 2003; 43(1):70–84. [PubMed: 14499463]

Lin D, Zhang J, et al. Group sparse canonical correlation analysis for genomic data integration. BMC Bioinform. 2013; 14:245.

Liu J, Huang J, et al. Incorporating group correlations in genome-wide association studies using smoothed group Lasso. Biostatistics. 2013; 14(2):205–219. [PubMed: 22988281]

Meda SA, Bhattarai M, et al. An fMRI study of working memory in first-degree unaffected relatives of schizophrenia patients. Schizophr. Res. 2008; 104(1–3):85–95. [PubMed: 18678469]

Ng B, Abugharbieh R. Generalized sparse regularization with application to fMRI brain decoding. Inform. Process. Med. Imag. 2011; 22:612–623.

Ohtsuki T, Sakurai K, et al. Mutation analysis of the NMDAR2B (GRIN2B) gene in schizophrenia. Mol. Psychiatr. 2001; 6(2):211–216.

Onitsuka T, Shenton ME, et al. Middle and inferior temporal gyrus gray matter volume abnormalities in chronic schizophrenia: an MRI study. Am. J. Psychiatr. 2004; 161(9):1603–1611. [PubMed: 15337650]

Parkhomenko E, Tritchler D, et al. Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. 2009; 8(1)

Pinault D. Dysfunctional thalamus-related networks in schizophrenia. Schizophr. Bull. 2011; 37(2): 238–243. [PubMed: 21307040]

Porteous DJ, Thomson P, et al. The genetics and biology of DISC1 – an emerging role in psychosis and cognition. Biol. Psychiatr. 2006; 60(2):123–131.

Posada D, Buckley TR. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 2004; 53(5): 793–808. [PubMed: 15545256]

Qin S, Zhao X, et al. An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray. Eur. J. Hum. Genet. 2005; 13(7):807–814. [PubMed: 15841096]

Shenton ME, Dickey CC, et al. A review of MRI findings in schizophrenia. Schizophr. Res. 2001; 49(1–2):1–52. [PubMed: 11343862]

Simon N, Tibshirani R. Standarization and the group lasso penalty. Stat. Sin. 2012; 22:983–1001.

Sui J, Pearlson G, et al. Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. Neuroimage. 2011; 57(3):839–855. [PubMed: 21640835]

Sutrala SR, Goossens D, et al. Gene copy number variation in schizophrenia. Schizophr. Res. 2007; 96(1–3):93–99. [PubMed: 17826036]

Szycik GR, Munte TF, et al. Audiovisual integration of speech is disturbed in schizophrenia: an fMRI study. Schizophr. Res. 2009; 110(1–3):111–118. [PubMed: 19303257]

Torrey EF. Schizophrenia and the inferior parietal lobule. Schizophr. Res. 2007; 97(1–3):215–225. [PubMed: 17851044]

Tyekucheva S, Marchionni L, et al. Integrating diverse genomic data using gene sets. Genome Biol. 2011; 12(10):R105. [PubMed: 22018358]

Tzourio-Mazoyer N, Landeau B, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002; 15(1):273–289. [PubMed: 11771995]

Waaijenborg S, Zwinderman AH. Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis. Bioinformatics. 2009; 25(21):2764–2771. [PubMed: 19689958]

Waaijenborg S, Hamer PCVDW, et al. Quantifying the association between gene expressions and DNA-Markers by penalized canonical correlation analysis. Stat. Appl. Genet. Mol. Biol. 2008; 7(1)

Wiley HS. Integrating multiple types of data for signaling research: challenges and opportunities. Sci. Signal. 2011; 4(160):pe9. [PubMed: 21325205]

Wilson GM, Flibotte S, et al. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. Hum. Mol. Genet. 2006; 15(5):743–749. [PubMed: 16434481]

Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol. 2009; 8(1)

Witten DM, Tibshirani R, et al. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–534. [PubMed: 19377034]

Wu TT, Chen YF, et al. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009; 25(6):714–721. [PubMed: 19176549]

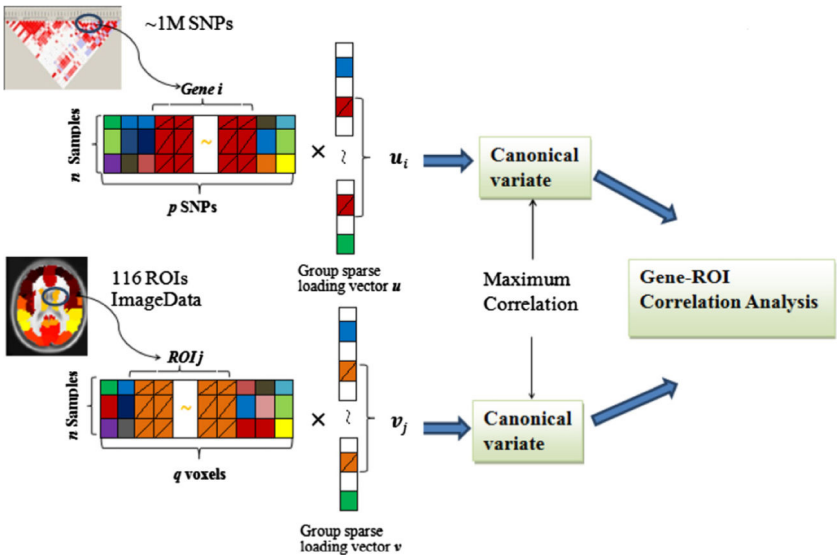

Fig. 1.

A schematic illustration of combining both fMRI image and SNP data by the group sparse CCA model to identify correlated genes and ROIs.

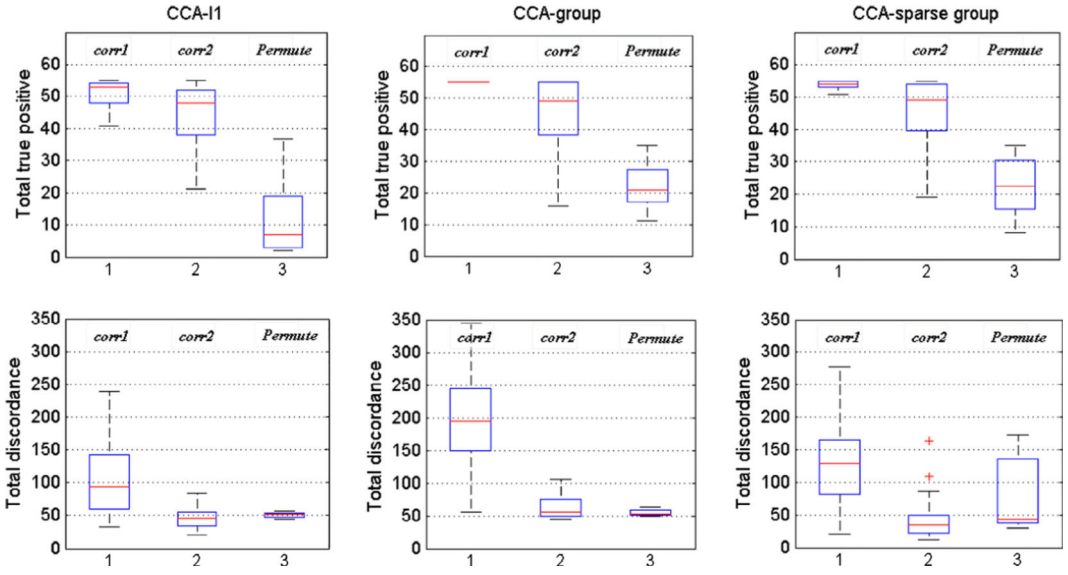

Fig. 2.

The total true positive number (TPP) and total discordance (TD) of estimating correlations using CCA-*l*1, CCA-group and CCA-sparse group methods using three different parameter selection criteria: corr1 (Eq. (10)) proposed by Parkhomenko et al. (2009), corr2 (Eq. (11)) proposed by Waaijenborg et al. (2008) and permutation based criterion (Eq. (12)) proposed by Witten and Tibshirani (2009).

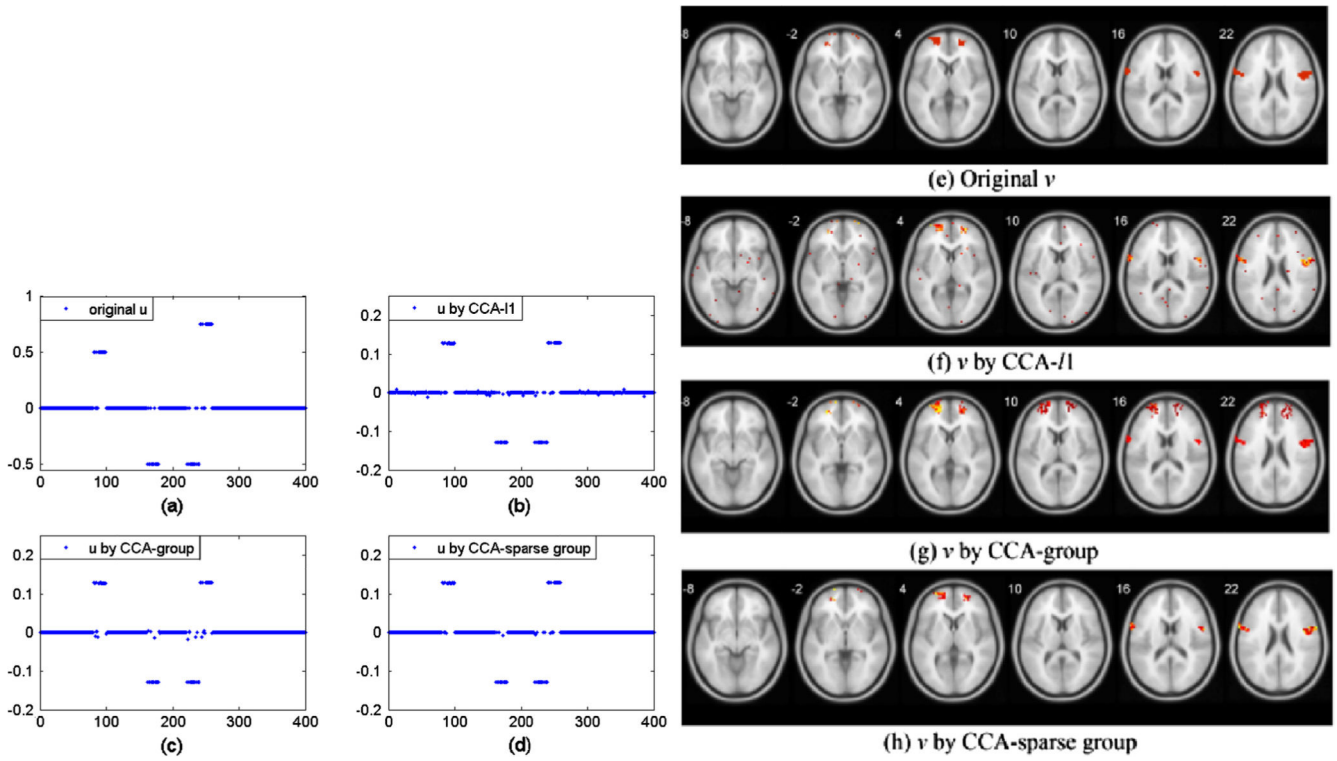

Fig. 3.

A comparison of the performance of group sparse CCA and three sCCA methods. (a) True ***u***; (b–d) ***u*** recovered by CCA-*l*1, CCA-group and CCA-sparse group respectively. (e) True ***v***; (f–h) ***v*** recovered by three sCCA methods.

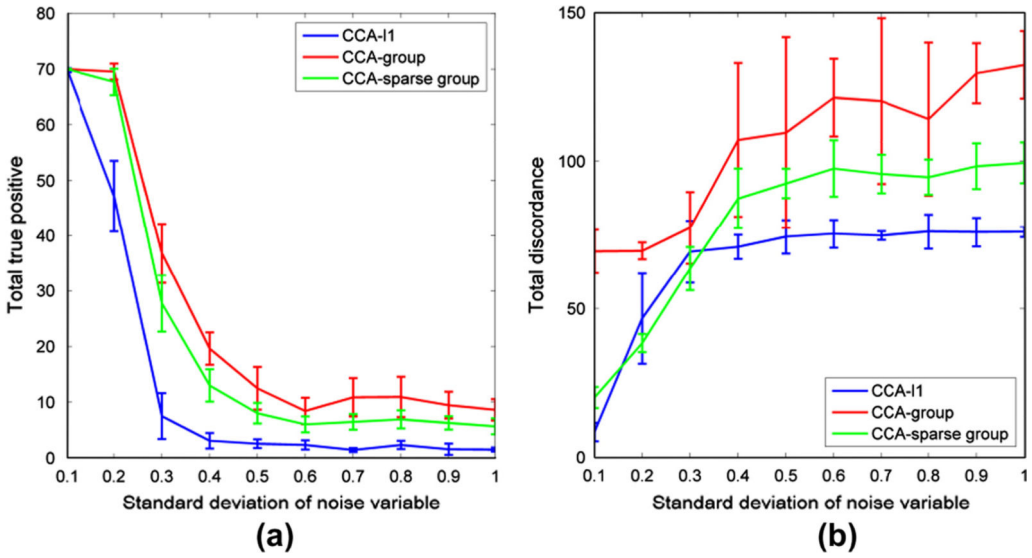

Fig. 4.

A comparison of three methods for different correlation level influenced by noise. (a) The value of total true positive obtained by three methods when the standard deviation of noise increases from 0.1 to 1, showing that the highest correlation of true variables between two data sets is within the range from 0.958 to 0.18. (b) The total discordance by three methods when the standard deviation of noise changes from 0.1 to 1.

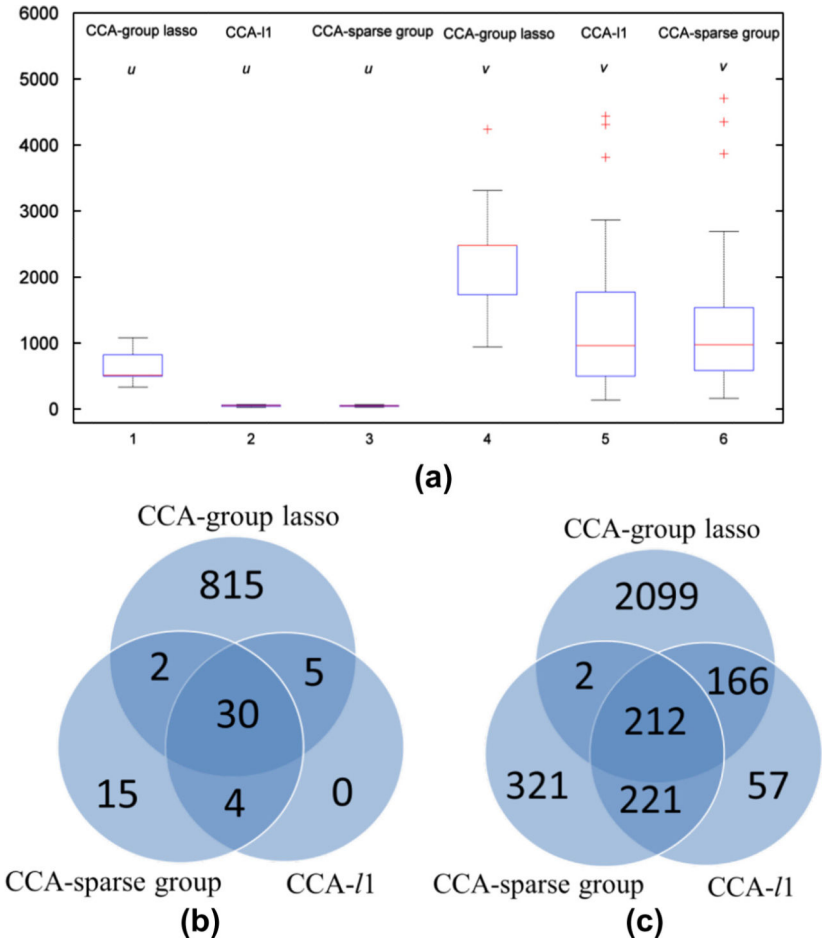

Fig. 5.

The comparison of the number of features in loading vectors ***u*** and ***v*** selected by three methods. (a) The number of SNPs and voxels selected by three methods with 50 sub-samplings. (b–c) Show the top ranked SNPs (b) and voxels (c) in the first pair of canonical variates with cut-off threshold = 0.3 selection probability.

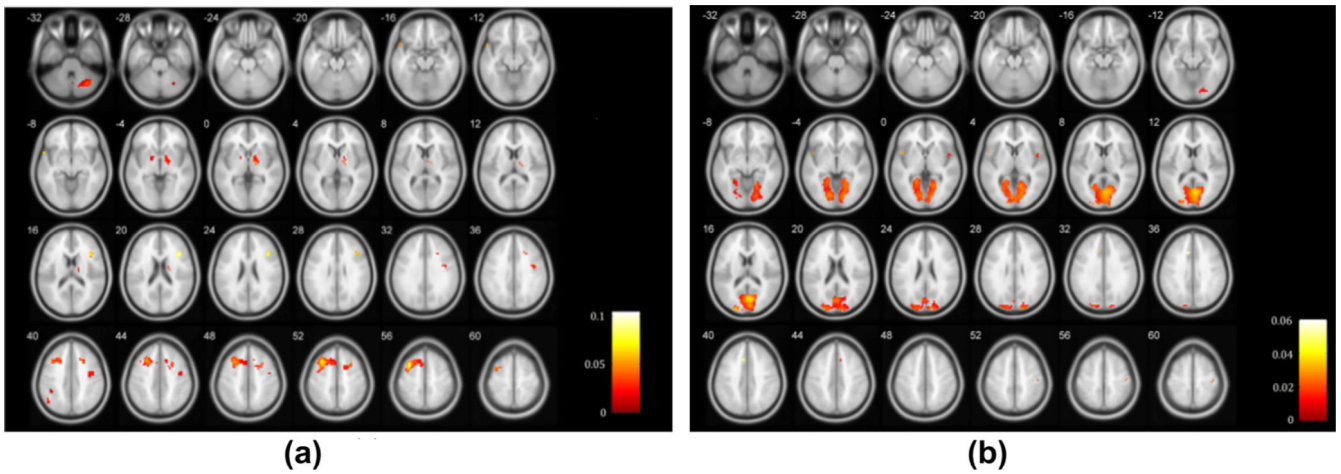

Fig. 6.

Maps showing regions correlated with genetic factors in the first (a) and second (b) canonical variates.

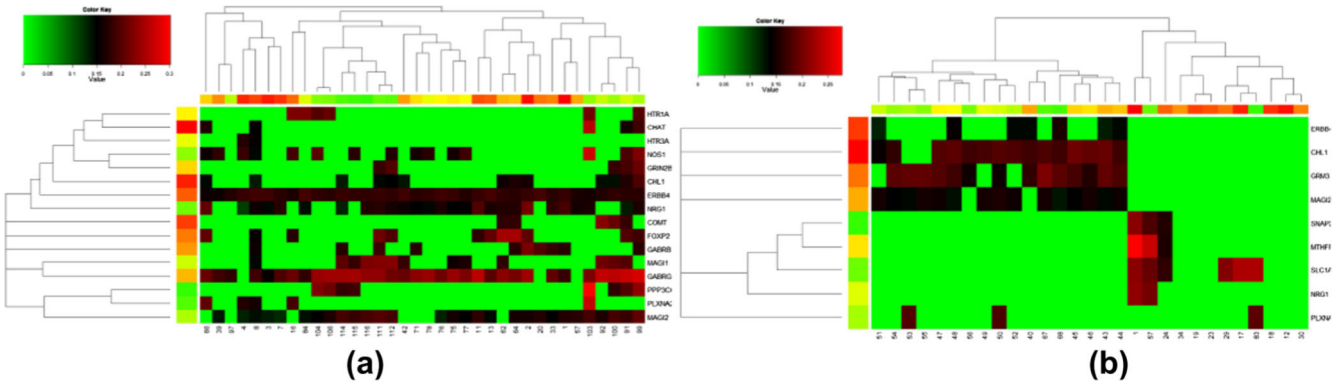

Fig. 7.

Heatmap of the gene-ROI correlations derived from the first (a) and second (b) pairs of canonical variates.

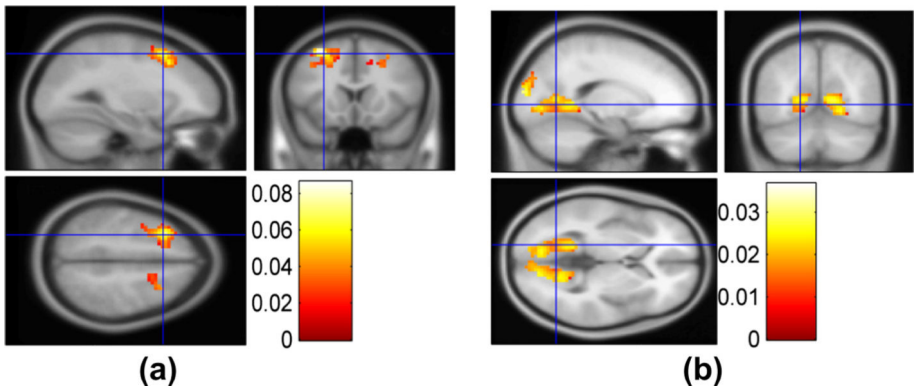

Fig. 8.

(a) ROIs 3, 4, 7, 8 which are identified to be mainly correlated with genes ERBB4, NRG1, MAGI1 and GABRG2. (b) ROIs 40, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 67, 68 which are found to correlate with Gene ERBB4, CHL1, GRM3 and MAGI2. Note: the index of ROI is given by the aal template (Tzourio-Mazoyer et al., 2002).

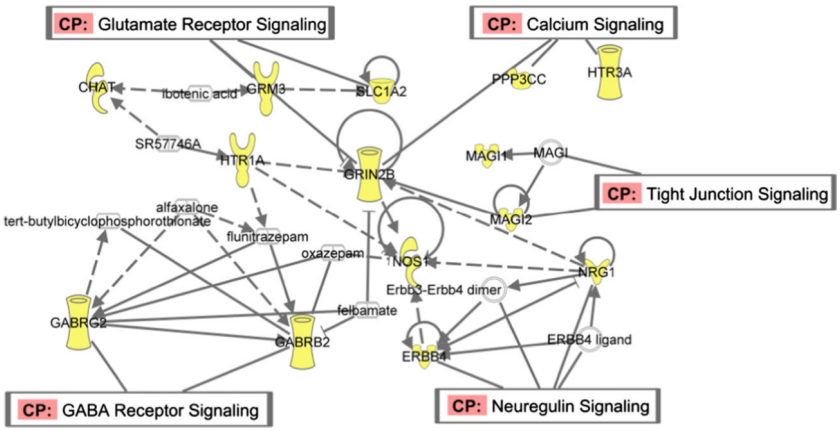

Fig. 9.

Functional network built based on the genes identified within the two pairs of canonical variates.

Table 1

The iterative algorithm of group sparse CCA.

1. Initialize $\boldsymbol{u}^0$ and $\boldsymbol{v}^0$ by traditional CCA decomposition, $\|\boldsymbol{u}^0\|_2 = 1$, $\|\boldsymbol{v}^0\|_2 = 1$.
2. Solve $\boldsymbol{U}^j, \boldsymbol{V}^j$ using the following iterations until it convergence:
   (a) Fix $v = v^{j-1}$, $u^j \leftarrow \arg\min_{u,d} \| K - duv^t \|_F^2 + \Psi(u) \quad s.t. \quad \|u\|_2^2 \le 1$
   (b) Fix $u = u^j$, $v^j \leftarrow \arg\min_{v,d} \| K - duv^t \|_F^2 + \Phi(v) \quad s.t. \quad \|v\|_2^2 \le 1$
   (c) $d^j \leftarrow tr(\boldsymbol{K}\boldsymbol{v}^j(\boldsymbol{u}^j)^t)$ or $tr(\boldsymbol{K}\boldsymbol{u}^j(\boldsymbol{v}^j)^t)$
3. Update the remaining matrix $\boldsymbol{K} \leftarrow \boldsymbol{K} - tr(\boldsymbol{K}\boldsymbol{v}\boldsymbol{u}^t)\boldsymbol{u}\boldsymbol{v}^t$; go to Step (1) to obtain the next pair of loading vectors ($\boldsymbol{u}, \boldsymbol{v}$).

Table 2

Block coordinate decent for group sparse CCA.

Input: iteration step j, $\boldsymbol{u}^j$, $\boldsymbol{v}^j$, $\|\boldsymbol{u}^j\|_2 = 1$, $\|\boldsymbol{v}^j\|_2 = 1$, λ_1 , τ_1.

Output: $\boldsymbol{u}^{j+1}$

Solve $\boldsymbol{u}^j$ using block coordinate decent until it convergence:

1. For each group $k = 0$ to L

2. $\text{Soft}_k(\boldsymbol{Kv}) = S((\boldsymbol{Kv})_k, \tau_1)$, *where* $S(\cdot)$ is the soft-thresholding function.

3. If $\| \ Soft_k(Kv) \|_2 \leq \lambda_1$ then $u_k^{j+1} = 0$.

4. Else

5. $Sg_k(Kv) = \frac{1}{2}\left[Soft_k(Kv) - \lambda_1 \omega_l \frac{Soft_k(Kv)}{\| \ Soft_k(Kv) \|_2}\right]$.

6. Update $u_k^{j+1} \frac{Sg_k(Kv)}{\Delta}$.

7. End if

8. End for

9. Update $\Delta = \| \ [Sg^1(Kv), Sg^2(Kv), \ldots, Sg^k(Kv)] \|$ to make $\| \ u^{j+1} \|_2^2 = 1$.

10. Repeat (1-9), until $\|\boldsymbol{u}^{j+1} - \boldsymbol{u}^j\|_2 \leq \varepsilon$, else $\boldsymbol{u}^j = \boldsymbol{u}^{j+1}$.

Table 3

The count of true positive (TP), false positive (FP) and discordance by three methods.

Model	TP (voxels/SNPs)	FP (voxels/SNPs)	Discordance (voxels/SNP)
CCA-/1	185/60	912/36	941/36
CCA-group	214/60	4097/20	4097/20
CCA-sparse group	200/60	113/4	127/4

Table 4

Correlation coefficients based on the variants estimated by three models in the test data.

	CCA-group lasso	CCA-*l*1	CCA-sparse group
Test sample correlation (mean ± SD)			
1st pair	0.4497±(0.043)	0.4235±(0.057)	0.4527±(0.048)
2nd pair	0.4137±(0.034)	0.3959±(0.045)	0.4292±(0.033)

Table 5

The list of genes selected in the first and second pair of canonical variates.

Canonical variate	Gene ID	SNPs number
1st pair	CHAT CHL1 COMT ERBB4 FOXP2 GABRB2 GABRG2 GRIN2B HTR1A HTR3A MAGI1 MAGI2 NOS1 NRG1 PLXNA2 PPP3CC	51
2nd pair	CHL1 ERBB4 GRM3 MAGI2 MTHFR NRG1 PLXNA2 SLC1A2 SNAP29	19

Table 6

The correlated brain ROIs selected in two pairs of canonical variates.

1st pair			2nd pair		
Brain region	**Brodmann area**	**L/R volume (cm^3)**	**Brain region**	**Brodmann area**	**L/R volume (cm^3)**
Superior Frontal Gyrus	6, 8	3.1/0.1	Lingual Gyrus	17, 18, 19	5.4/6.5
Inferior Parietal Lobule	40	0.6/*	Posterior Cingulate	18, 23, 30, 31	2.3/2.9
Medial Frontal Gyrus	6, 8, 32	0.7/0.3	Parahippocampa Gyrus	18, 19, 30	1.2/0.4
Precentral Gyrus	6	0.3/1.2	Precuneus	19, 31	0.7/0.8
Middle Frontal Gyrus	6, 8	4.5/0.6	Superior Temporal Gyrus	22	*/0.1
Thalamus	*	0.1/0.4	Cingulate Gyrus	32	0.5/0.2
Cingulate Gyrus	6	0.3/1.2	Cuneus	17, 18, 19, 23, 30	8.1/9.3
Culmen	*	0.1/0.5	Middle Occipital Gyrus	18	0.8/0.1
Declive	*	0.2/0.6	Insula	13	0.1/0.2