INTEGRATION OF MULTIPLE GENOMIC IMAGING DATA FOR THE STUDY OF SCHIZOPHRENIA USING JOINT NONNEGATIVE MATRIX FACTORIZATION

Min Wang[1,2], *Ting-Zhu Huang*[1], *Vince D.Calhoun*[3], *Jian Fang*[2], *Yu-Ping Wang*[2*]

[1]*School of Mathematical Sciences/Research Center for Image and Vision Computing, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China*

[2]*Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118, USA*

[3]*The Mind Research Network, University of New Mexico, NM 87131, USA*

ABSTRACT

Schizophrenia (SZ) is a complex disease caused by a lot genetic variants, epigenetic and brain region abnormalities. In this study, we adopted a joint nonnegative matrix factorization method to integrate three datasets including single nucleotide polymorphism (SNP), brain activity measured by functional magnetic resonance imaging (fMRI) and DNA Methylation to identify multi-dimensional modules associated with SZ. They are then used to study the coordination between regulatory mechanisms at multiple levels. This method projects multiple types of data onto a common feature space, in which heterogeneous variables with large coefficients on the same projected bases form a multi-dimensional module. The genomic factors in such modules have significant correlations and likely functional associations with brain activities. We applied this method to the real data analysis and identified multi-dimensional modules including SNP, fMRI and DNA methylation sites. These selected biomarkers were finally used to identify genes and voxels, which were confirmed to be significantly associated with SZ.

Index Terms— Joint nonnegative matrix factorization, SNP, fMRI, Methylation, Feature selection

1. INTRODUCTION

Genetic variations are recognized as important factors for schizophrenia (SZ). Recent years have seen many research works on exploring critical genes associated with SZ. Many potential genetic variants have also been identified as possible risk factors, for example, the G72/G30 gene locus on chromosome 13q [1], gene DISC1 variation [2] and copy number variations on gene GRIK3, EFNA5, AKAP5 and CACNG2 [3, 4]. In addition, DNA methylation as one of the main epigenetic mechanisms to regulate gene expression, has also been determined to be involved with the development of SZ. Davies *et al* [5] have found that the interindividual variations of DNA methylation are significantly correlated across blood and brain. Some studies have used blood DNA methylation to identify potential biomarkers for SZ disease status [6, 7]. In addition to (epi)-genetic approaches, fMRI has also been used by measuring brain activity and further identifying functional abnormalities within brain regions of SZ patients [8].

Different Datasets (e.g. SNP, fMRI and methylation data) represent the same biological samples in different views. These datasets provide partial and complementary information, and their joint analysis has the potential to reveal factors underlying complex diseases. Since different types of genomic data have different scales and formats, we cannot simply aggregate them for joint analysis. Various data fusion methods have been developed to address this challenge. In SZ study, most of the works either use single dataset [9, 10, 11] or two datasets [12]. However, there is little work that can take advantage of three or more datasets to achieve a more comprehensive analysis. Therefore, we propose to use nonnegative matrix factorization (NMF) [13] method for multiple data integration. NMF factorizes data matrix into parts-based representation with nonnegative constraints, which was widely used in data integration [14]. Zhang *et al* [15] proposed a joint NMF framework for pattern discovery from cancer genomic data. In this study, we employ joint NMF model to extract correlative modules across SNP, fMRI and methylation patterns for SZ. The correlative modules were then applied to identify significant genes or biomarkers associated with SZ.

The rest of the paper is organized as follows. In Section 2, we introduce the concept of joint NMF model and an optimization algorithm for solving the model. In Section 3 we present the results of real data analysis. Finally, in Section 4 we conclude the paper and discuss future directions.

2. MATERIALS AND METHODS

2.1. Data preparation and preprocessing

Participants in this study were from the Mind Clinical Imaging Consortium (MCIC). 80 SZ patients (age: 34 ± 11, 20 females) and 104 healthy controls (age: 32 ± 11, 38 females)

*Corresponding author. E-mail: wyp@tulane.edu

were analyzed here. We used three types of data (SNP, fMRI, DNA methylation data) shared by the 184 samples. We followed the same preprocessing procedures in [12] for SNP, fMRI data and [11] for methylation, resulting in 722,177 SNPs, 41,236 voxels and 27,508 methylation sites. Since we want to find the biomarkers only associated with SZ, we applied t-test on these three datasets between SZ and healthy samples and only selected those variables with P-value< 0.05). After variable selection, we obtained 50,452 SNPs, 2,550 voxels and 2724 methylation sites across 184 samples, which were represented in three matrices of size 184×50452, 184×2550 and 184×2724, respectively.

We standardized each column of the three matrices and scaled the matrices so that they have the same Frobenius norm. We employed the method used in [16] to make the matrices nonnegative. Specifically, each column in the matrix was represented with two columns. The first column stores the positive values and the second column stores the absolute value of the negative values. The rest of the matrix were filled with zeros.

2.2. The joint NMF model

NMF is a powerful tool for data reduction, which has been used in analyzing high-throughput genomic data [16]. For a nonnegative data matrix $X \in \mathbb{R}^{m\times n}$, the NMF can be formulated as

$$\begin{aligned} &\min_{W,H} \|X - WH\|_F^2 \\ &s.t. \;\; W \geq 0, H \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $W \in \mathbb{R}^{m\times r}$ stores the basis column vectors and $H \in \mathbb{R}^{r\times n}$ stores the corresponding column coefficient vectors. r is the number of the basis vectors. In most cases, $r < \min(m, n)$.

For each column $x_{\cdot j}$, a linear, nonnegative approximation of the basis vectors is given by

$$x_{\cdot j} = \sum_{i=1}^{r} w_{\cdot i} h_{i,j} = W h_{\cdot j}. \tag{2}$$

From Eq. (2) we can know that the r basis vectors $w_{\cdot i}$ can be considered as the skeleton of the data, and the r-dimensional coefficient vector $h_{\cdot j}$ describes the weights of each skeleton on $x_{\cdot j}$.

For the SNP, fMRI and DNA methylation datasets of the same samples, we denote them by $X_1 \in \mathbb{R}^{m\times n_1}$, $X_2 \in \mathbb{R}^{m\times n_2}$, $X_3 \in \mathbb{R}^{m\times n_3}$, respectively. Here m is the number of samples we used. n_1, n_2 and n_3 are the number of SNPs, voxels and methylation sites, respectively. To simplify the description, we assume the three matrices to be nonnegative. We jointly decompose the three data matrices into a common basis matrix $W \in \mathbb{R}^{m\times r}$ with different coefficient matrices $H_q \in \mathbb{R}^{r\times n_q}, (q = 1, 2, 3)$ and this process is called joint NMF, formulated as follows:

$$\begin{aligned} &\min_{W,H_1,H_2,H_3} \sum_{q=1}^{3} \|X_q - WH_q\|_F^2 \\ &s.t. \;\; W \geq 0, H_q \geq 0, q = 1, 2, 3. \end{aligned} \tag{3}$$

We use the multiplicative algorithm introduced in [15] to find optimal solution to the problem in Eq. (3). We randomly initialize matrices W and H_1, H_2 and H_3 with nonnegative values and update them iteratively by using the generalized multiplicative updata rules as follows

$$W_{ij} = W_{ij} \frac{(X_1H_1^T + X_2H_2^T + X_3H_3^T)_{ij}}{(W(X_1H_1^T + X_2H_2^T + X_3H_3^T))_{ij}}, \tag{4}$$

$$(H_q)_{ij} = (H_q)_{ij} \frac{(W^T X_q)_{ij}}{(W^T W H_q)_{ij}}, q = 1, 2, 3. \tag{5}$$

The iterations are terminated until the relative value of the residue of the object function in Eq. (3) is smaller than a predefined tolerance τ. Since the object function is nonconvex on both W and H, the above algorithm may only find a local minimum. We rerun the procedure for 100 times with different initial values. The solution with the lowest value of object function was used as the final solution for further analysis.

2.3. Identification of modules

After joint NMF, the three datasets are projected onto a common space whose basis vectors are stored in the factor matrix W. The coefficient row vectors in matrices H_1, H_2 and H_3 are used to identify memberships of SNPs, voxels and methylation sites in each modules, respectively. We calculate the z-score for each element in each row of $H_q(q = 1, 2, 3)$ by

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{6}$$

where μ_i is the mean value of i-th row vector in $H_q(q = 1, 2, 3)$ and σ_i is the standard deviation. For H_q, $z_{ij} > T$ means that the feature j in dataset X_q is a member of module i, where $q = 1, 2, 3$ and $T > 0$ is a given threshold.

Since the correlation between imaging and genetic data could be quite low and the number of variables is much larger than the number of samples, the above module identification procedure could still yield many irrelevant features. As an alternative, we apply stability selection [17] to extract correlated features in each datasets. We perform random sampling of size 92 from original 184 samples for 1000 times and employ joint NMF on each subsample to find the r modules. For each subsample, we concatenate the feature vectors in r modules to be one feature vector. In this way, each subsample set of size 92 yielded three feature vectors corresponding to SNPs, fMRI, methylation sites, respectively. For each type of datasets, we rank the feature indices in all of those 1000 feature vectors according to their occurrence frequency. We select the features whose occurrence probability are larger than

θ_1, θ_2 and θ_3 as the significant biomarkers corresponding to SNP, fMRI, methylation, respectively. Finally, we perform joint NMF on all of the 184 samples and use the intersection of the resulting r modules from the three feature vectors as final modules. If there exist modules with null feature indices in any datasets, it indicates that the correlation may not exist across three datasets.

2.4. Significance estimation

We employ a permutation test to estimate the significance (P-value) of the identified modules. Assuming the number of SNPs, voxels and methylation sites in a module are l_1, l_2 and l_3, we denote them by SNPs: $A = [a_1, a_2, \ldots, a_{l_1}]$, fMRI voxels: $B = [b_1, b_2, \ldots, b_{l_2}]$, methylation: $C = [c_1, c_2, \ldots, c_{l_3}]$, respectively. Note that a_i, b_j, c_k are all vectors and the length of vector is the number of samples. We use $\rho(x, y)$ to represent the Pearson correlation between x and y. Based on the above assumptions, the mean correlation among the three datasets in a module can be given as

$$\rho^* = \frac{1}{3}\left(\frac{1}{l_1 l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}\rho(a_i, b_j) + \frac{1}{l_1 l_3}\sum_{i=1}^{l_1}\sum_{k=1}^{l_3}\rho(a_i, c_k) + \frac{1}{l_2 l_3}\sum_{j=1}^{l_2}\sum_{k=1}^{l_3}\rho(b_j, c_k)\right). \tag{7}$$

We permutate the row order of matrices A and B while keep matrix C unchanged for φ times. For each permutation, the mean correlation ρ_i^* $(i = 1, 2, \ldots, \varphi)$ can be calculated by Eq. (7), which is to build the null distribution of the mean correlation. By large number of permutations, the significance of the mean correlation can be evaluated by

$$\text{P-value} = |\{i|\rho_i^* \geq \rho^*, i = 1, 2, \ldots, \varphi\}|/\varphi, \tag{8}$$

where $|S|$ is the number of elements in set S. Variables with P-value smaller than 0.01 were considered to be significant.

3. RESULTS

3.1. Parameter selection

There are six parameters to be determined before we perform the method introduced in Section 2 on the SNP, fMRI and DNA methylation datasets (184 samples). We found that when $r > 4$, the modules we identified have at least one null module, which indicates that the datasets were over-factorized. Therefore, we set $r = 4$. T used in preliminary selection was set to $T = 3$ which is a z-score that corresponds to the P-value smaller than 0.01. θ_1,θ_2 and θ_3 were used to select the significant SNPs, fMRI voxels and methylations based on the occurrence probability in the stability selection procedure. If we set a larger $\theta_i, (i = 1, 2, 3)$, we will select less variables (SNPs, voxels and methylations) in the identified modules. Moreover, we use cross validation adopting a criterion that maximizes the mean correlation of the modules to select the parameters $\theta_i, (i = 1, 2, 3)$ from $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ with different combinations. Based on the above principles, we found that $\theta_1 = \theta_3 = 0.7$, $\theta_2 = 0.4$ perform better. In a word, we set $r = 4$, τ =1.e-6, $T = 3$, $\theta_1 = 0.7$, $\theta_2 = 0.4$, $\theta_1 = 0.7$ in our test.

Table 1. The list of genes corresponding to SNP and methylation data.

Dataset	Gene ID
SNP	ABCA3, APPBP2, C17orf64, CALB1, CAP2, DOK5, KCNK3, LOC645101, LOC729602, METT5D1, MYO16, SNX20
Methylation	ANKRD15, ATP6V0A4, C1orf116, C1orf172, C21orf56, CCNA1, CREB3L3, FLJ11017, HOXA4, HTN3, KIF27, LDHD, PAGE5, PDIA2, RCBTB2, RPL26L1, TMEM100

Table 2. The brain regions corresponding to fMRI data.

Brain region	L/R volume(cm^3)
midcingulate area	0.56/0.86
parahippocampal gyrus	*/0.405
postcentral gyrus	1.701/*
supramarginal gyrus	*/0.459
angular gyrus	*/0.27
precuneus	0.162/0.675
superior temporal gyrus	0.297/*

3.2. Module analysis

Within the 4 modules we identified, only one module has three feature indices corresponding to SNP, fMRI and methylation. There are 13 SNPs from 12 genes, 18 methylations from 17 genes and 210 voxels from 9 brain regions selected in the module. The list of genes and brain regions were presented in Table. 1 and Table. 2, respectively. Moreover, the selected voxels were plotted in Fig. 1. Since the three datasets all have some feature indices in the module, we calculate the mean correlation and P-value not only among the three datasets by Eq. (7),(8), but also between any two datasets in the cases and controls, respectively. The correlation and P-value were displayed in Table. 3. The correlations of SNP-fMRI and fMRI-methylation are 0.204 and 0.124 in cases, which are much higher than in controls. The P-value of SNP-fMRI and fMRI-methylation in cases are all smaller than 0.01, which means the correlations are significant. But the P-value of SNP-fMRI and fMRI-methylation in controls are 0.281 and 0.991 , which are all larger than 0.01. In SNP-methylation and SNP-fMRI-methylation, the correlations in

Table 3. The correlation and P-value of the module in cases and controls.

Dataset combination	Cases	Controls
	Correlation(P-value)	
SNP-fMRI	0.204(1.2e-3)	0.086(0.281)
SNP-methylation	0.378(1.3e-4)	0.213(6.4e-4)
fMRI-methylation	0.124(7.2e-3)	0.064(0.991)
SNP-fMRI-methylation	0.235(1.0e-4)	0.121(1.2e-3)

cases are also much higher than in controls and the P-value are all smaller than 0.01, so the P-values in cases are more significant. The results in Table. 3 further confirm that the SNPs, fMRI voxels and methylation sites in the identified module are SZ-specific, which can be used to next study the genes and brain regions associated with SZ.

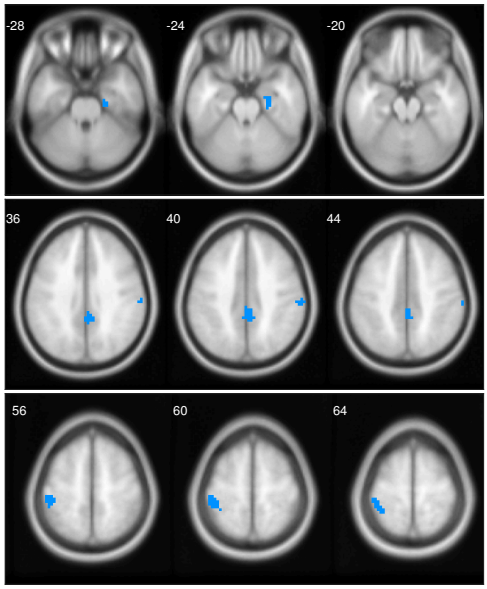

Fig. 1. Maps showing regions correlated with SNPs and methylation datasets.

There are 29 genes identified from SNP and Methylation data. Among them, the gene APPBP2 encoding amyloid proteins were reported to have potential relationship with the SZ [18]. Gene CAP2 is a central nervous system-expressed functional gene and it has a high rate of copy number variation in SZ [19]. C21orf56 is a up-regulated gene differentially expressed between SZ cases and controls, which has potential relations with SZ [20]. CCNA1 is very important in cell cycle regulation, which is significantly dysregulated in SZ cases [21]. FLJ11017 is reported to have a role in SZ [22] and KIF27 is reported to be differentially expressed in bipolar disorder and/or SZ [23].

We identified 9 brain regions from the fMRI data. In particular, for the left superior temporal gyrus, patients with first-episode SZ showed significant decreases in gray matter volume over time [24]. The left postcentral gyrus was reported to show significant difference between the SZ cases and controls [25]. The regional heterogeneity of midcingulate area folding complexity may be related to SZ with altered cortical developmental pathology [26]. In [27], an abnormally increased parahippocampal response to neutral faces was positively associated with reality distortion in SZ.

4. CONCLUSIONS

SNP, fMRI and methylation are all important factors to study the SZ, but most of the existing approaches either focus on one or two datasets analysis. If we represent the datasets as matrices with identical rows for the same sample sets, joint NMF simultaneously projects matrices onto a lower dimension space and represents each column in the matrices with a linear, nonnegative approximation of the primary bases. For each matrix, we can use the nonnegative coefficient values to select the columns correlated with each basis. In this way, the hidden dependence structures can be identified and the data heterogeneity in the datasets will also be reduced. Joint NMF can be extended to handle three or more matrices easily. In this paper, we employ a joint NMF model to extract important modules correlated across three datasets and identify SZ-specific features, which can be used to better understand the biological mechanisms underlying brain functions. Since joint NMF model does not use any prior information, in future studies, we will incorporate gene networks or brain region network information into our analysis model.

Acknowledgement

This work is supported by 973 Program (Grant No. 2013CB329404), NSFC (Grant No. 61370147), the Fundamental Research Funds for the Central Universities (ZYGX2013Z005), NIH (R01 MH104680, R01 MH107354, R01 GM109068)

5. REFERENCES

[1] J.A. Badner et al., "Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia," *Molecular psychiatry*, vol. 7, no. 4, pp. 405–411, 2002.

[2] J.H. Callicott et al., "Variation in disc1 affects hippocampal structure and function and increases risk for schizophrenia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 24, pp. 8627–8632, 2005.

[3] G.M. Wilson et al., "Dna copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling," *Human molecular genetics*, vol. 15, no. 5, pp. 743–749, 2006.

[4] S.R. Sutrala et al., "Gene copy number variation in schizophrenia," *Schizophrenia research*, vol. 96, no. 1, pp. 93–99, 2007.

[5] M.N. Davies et al., "Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood," *Genome biology*, vol. 13, no. 6, pp. 1, 2012.

[6] E.L. Dempster et al., "Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder," *Human molecular genetics*, p. ddr416, 2011.

[7] Y.B. Chen et al., "Effects of maoa promoter methylation on susceptibility to paranoid schizophrenia," *Human genetics*, vol. 131, no. 7, pp. 1081–1087, 2012.

[8] G.R. Szycik et al., "Audiovisual integration of speech is disturbed in schizophrenia: an fmri study," *Schizophrenia research*, vol. 110, no. 1, pp. 111–118, 2009.

[9] R.E. Gur et al., "An fmri study of facial emotion processing in patients with schizophrenia," *American Journal of Psychiatry*, vol. 159, no. 12, pp. 1992–1999, 2002.

[10] T. Li et al., "Family-based linkage disequilibrium mapping using snp marker haplotypes: application to a potential locus for schizophrenia at chromosome 22 q 11," *Molecular psychiatry*, vol. 5, no. 1, pp. 77–84, 2000.

[11] J.Y. Liu et al., "Methylation patterns in whole blood correlate with symptoms in schizophrenia patients," *Schizophrenia bulletin*, vol. 40, no. 4, pp. 769–776, 2014.

[12] D.D. Lin et al., "Correspondence between fmri and snp data by group sparse canonical correlation analysis," *Medical image analysis*, vol. 18, no. 6, pp. 891–902, 2014.

[13] D.D Lee et al., "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[14] Ariana Anderson et al., "Non-negative matrix factorization of multimodal mri, fmri and phenotypic data reveals differential changes in default mode subnetworks in adhd," *NeuroImage*, vol. 102, pp. 207–219, 2014.

[15] S.H. Zhang et al., "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic acids research*, p. gks725, 2012.

[16] P.M. Kim et al., "Subsystem identification through dimensionality reduction of large-scale gene expression data," *Genome research*, vol. 13, no. 7, pp. 1706–1718, 2003.

[17] N. Meinshausen et al., "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[18] M.R. Barnes et al., "Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia," *Journal of neuroscience research*, vol. 89, no. 8, pp. 1218–1227, 2011.

[19] D. Malhotra et al., "High frequencies of de novo cnvs in bipolar disorder and schizophrenia," *Neuron*, vol. 72, no. 6, pp. 951–963, 2011.

[20] L. Sun et al., "Gene expression profiling in peripheral blood mononuclear cells of early-onset schizophrenia," *Genomics data*, vol. 5, pp. 169–170, 2015.

[21] Y. Wang et al., "Do shared mechanisms underlying cell cycle regulation and synaptic plasticity underlie the reduced incidence of cancer in schizophrenia?," *Schizophrenia research*, vol. 130, no. 1, pp. 282–284, 2011.

[22] T. Lencz et al., "Genetic prediction of schizophrenia susceptibility," Nov. 25 2010, US Patent App. 12/311,897.

[23] L. Shao, "Genes differentially expressed in bipolar disorder and/or schizophrenia," Mar. 27 2008, US Patent App. 11/712,827.

[24] K. Kasai et al., "Progressive decrease of left superior temporal gyrus gray matter volume in patients with first-episode schizophrenia," *American Journal of Psychiatry*, vol. 160, no. 1, pp. 156–164, 2003.

[25] D.E. Job et al., "Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry," *Neuroimage*, vol. 17, no. 2, pp. 880–889, 2002.

[26] I. Nenadic et al., "Cortical surface complexity in frontal and temporal areas varies across subgroups of schizophrenia," *Human brain mapping*, vol. 35, no. 4, pp. 1691–1699, 2014.

[27] S. Surguladze et al., "A reversal of the normal pattern of parahippocampal response to neutral and fearful faces is associated with reality distortion in schizophrenia," *Biological psychiatry*, vol. 60, no. 5, pp. 423–431, 2006.