Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 783592, 10 pages
http://dx.doi.org/10.1155/2015/783592

Research Article

Multiblock Discriminant Analysis for Integrative Genomic Study

Mingon Kang,[1] Dong-Chul Kim,[2] Chunyu Liu,[3] and Jean Gao[1]

[1] *Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA*
[2] *Department of Computer Science, University of Texas-Pan American, Edinburg, TX 78539, USA*
[3] *Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 66012, USA*

Correspondence should be addressed to Jean Gao; gao@uta.edu

Received 9 December 2014; Accepted 21 April 2015

Academic Editor: Klaus Wimmers

Human diseases are abnormal medical conditions in which multiple biological components are complicatedly involved. Nevertheless, most contributions of research have been made with a single type of genetic data such as Single Nucleotide Polymorphism (SNP) or Copy Number Variation (CNV). Furthermore, epigenetic modifications and transcriptional regulations have to be considered to fully exploit the knowledge of the complex human diseases as well as the genomic variants. We call the collection of the multiple heterogeneous data "multiblock data." In this paper, we propose a novel Multiblock Discriminant Analysis (MultiDA) method that provides a new integrative genomic model for the multiblock analysis and an efficient algorithm for discriminant analysis. The integrative genomic model is built by exploiting the representative genomic data including SNP, CNV, DNA methylation, and gene expression. The efficient algorithm for the discriminant analysis identifies discriminative factors of the multiblock data. The discriminant analysis is essential to discover biomarkers in computational biology. The performance of the proposed MultiDA was assessed by intensive simulation experiments, where the outstanding performance comparing the related methods was reported. As a target application, we applied MultiDA to human brain data of psychiatric disorders. The findings and gene regulatory network derived from the experiment are discussed.

1. Introduction

Human diseases involve complex processes that include interactive actions of biological multiple layers such as genetic, epigenetic, and transcriptional regulation. Conducting research based on a single type of biological data produces insufficient results to fully exploit the knowledge of the complex human diseases. The prior research shows that it is essential for the study to be based on a comprehensive consideration of the multiple biological data to grasp an in-depth understanding of the complex mechanisms of the human diseases and the identification of disease markers. The recent advances of high-throughput technologies such as DNA microarray and sequencing technologies efficiently profile various types of genomic data. The genomic data include Single Nucleotide Polymorphism (SNP), Copy Number Variation (CNV), DNA methylation (DM), and gene expression (GE). Integrative genomic analysis of the heterogeneous genomic data plays an important role in profiling a global view of a biological system as well as identifying significant markers of the human diseases.

However, most research has focused solely on investigations of a single type of the genomic data. Genome-Wide Association Studies (GWAS) examine genetic loci which are associated with a trait (e.g., major diseases) using the SNP data [1, 2]. GWAS normally compare the SNP arrays of two groups, disease (case) and normal (control) samples. If a genetic variation on a locus with the disease samples is statistically significant to the controls, the SNP is considered associated with the disease, whereas expression Quantitative Trait Loci (eQTL) studies have been actively done to identify genetic loci that regulate gene expression [3]. Combining the gene microarray data with GWAS not only enables the capture of gene regulatory interactions but also provides insight into the genetic mechanism that regulates gene expression variations. However, both GWAS and eQTL mapping studies still remain as a "*missing heritability*" problem [4].

In addition to SNP, Copy Number Variation (CNV) and DNA methylation (DM) have also been highlighted as key factors that affect the gene expression regulation. CNV is a structural alternation of DNA in which specific regions of the genome are deleted or duplicated on chromosomes. Although CNV is frequently observed even in healthy individuals, it is hypothesized that the variants may cause diseases by directly affecting gene dosage and gene expression [5, 6]. Specifically, whole-genome association studies of the relationship between CNV and diseases reported that gene expression levels in CNV regions are strongly related to the deletion or duplication of the regions [6]. Typically, the deletion of either particular regions within a gene or regulatory regions of a gene may result in a lower gene expression than what is normally expressed. DM is an epigenetic modification that occurred by the addition of methyl group to the cytosine or adenine of DNA. DM inhibits transcription of the genes with high levels of 5-methylcytosine in their promoter region or recruits proteins such as histone deacetylases that can modify histones [7, 8]. The functionality of DM consequently changes the gene expression levels even on the same DNA bases.

Thus, recent research has actively extended GWAS and eQTL mapping studies to the integrative association studies with multiple types of genomic data. Most integrative genomic research focuses on identifying genetic, epigenetic, or posttranscriptional factors that control gene expression regulation (or microRNA) by considering the complex interactions of SNP, CNV, and DM [9–11]. Specifically, the Cancer Genomic Atlas [9] conducted large-scale multidimensional analysis with SNP, CNV, DM, and GE to provide comprehensive genomic characterizations for brain cancer. In Aure et al.'s work [10], the combination effects of CNV and DM were examined to identify the association with alterations of miRNA expression in breast tumors. Wagner et al. [11] studied the relationship between SNP, DM, and GE via multiple eQTL analysis.

Most of the integration approaches have used step-by-step processes. Ordinarily, approaches filter candidate markers by using statistical techniques at the first step and find the final markers that satisfy certain criteria at the remaining stages [12–15]. This type of integration method often makes increased "*type II errors*" at each step, that is, fails to find informative markers by incorrectly identifying them as insignificant. Moreover, they do not consider interaction effects of the multiblock data. Mechanism was not considered.

Hence, research has recently started to shift toward approaches using systematical models in order to integrate and analyze the heterogeneous data comprehensively rather than through simple step-wise processes [16–18]. Multiblock methods of Partial Least Squares (PLS) and Generalized Canonical Correlation Analysis (GCCA) are representative methods. A derivative of a sparse version of PLS was proposed by penalizing both features and sample dimensions to identify "*regulatory modules*" [16]. Such PLS-based methods, which maximize the covariance between latent variables, often fail to detect significant factors when their intensities are weak. Furthermore, the method lacks the consideration of the discriminant analysis of the disease. A sparse multiblock analysis method derived from Generalized Canonical Correlation (SGCCA) was developed to identify multiblock association models while considering the relationship between the different data block such as *cis*-regulated mutations [17]. This work builds a hybrid model by combining both GWAS and eQTL models rather than a multiblock integration model. The data integration approach was suggested by utilizing multiple feature selection methods such as Principal Component Analysis (PCA), PLS, and LASSO [18]. They extracted the important factors using the dimensional reduction and feature selection methods and applied them on Cox survival models. However, combination effects of the multiblock data were ignored in this approach.

To tackle these limitations, we propose a novel Multiblock Discriminant Analysis (MultiDA) method for the integrative genomic study. The proposed method MultiDA makes the following main contributions.

(i) A new integrative genomic model for the discriminant analysis is introduced by exploiting class information.

(ii) A sophisticated optimal solution is developed to solve the discriminant analysis problem in the integrative genomic model.

First, we built a novel integrative genomic model for the discriminant analysis. The class data is considered as one block, and the total squared correlation including the class block is maximized. The introduction of the class block to the multiblock model enables us to perform discriminant analysis in the integrative genomic model. Secondly, we propose a sophisticated method to solve the discriminant analysis problem in the new integrative genomic model. The discriminant analysis is essential in identifying biomarkers of human diseases in computational biology. Regardless, it has been overlooked in the multiblock analysis. The efficient algorithm for the discriminant analysis and assessment of its performance are explored in this paper.

2. Methods

2.1. Notation. We suppose that there are J multiblock data. The multiblock data are measured on N numbers of the same set of observations. A block consists of a group of features that share common properties or represent one aspect of the sample. The multiblock data is denoted by $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_J\}$. The jth block data $\mathbf{X}_j$ is P_j-dimensional zero mean column vectors $\mathbf{X}_j \in \mathfrak{R}^{N \times P_j}$. A matrix $\mathbf{C} = \{c_{jk} \mid c_{jk} \in \{0, 1\}, 1 \leq j, k \leq J\}$ is a binary matrix that determines the linkage between the multiblock, where $c_{jk} = 1$ if the block j and the block k are connected or 0 if otherwise. In the proposed integrative genomic model, SNP, CNV, DM, GE, and class label (case or control) of the samples are considered as the multiblock components. For simplicity, $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, and $\mathbf{X}_5$ represent SNP, CNV, DM, GE, and class label, respectively. Through this paper, we use i for the index of the sample and $\{j, k\}$ for the multiblock. $(\imath)$ is used to denote a column vector of a matrix or an element of a vector. For instance, $\mathbf{X}_{i(\imath)}$ and $a_{i(\imath)}$ represent the $\imath$th column vector of the matrix

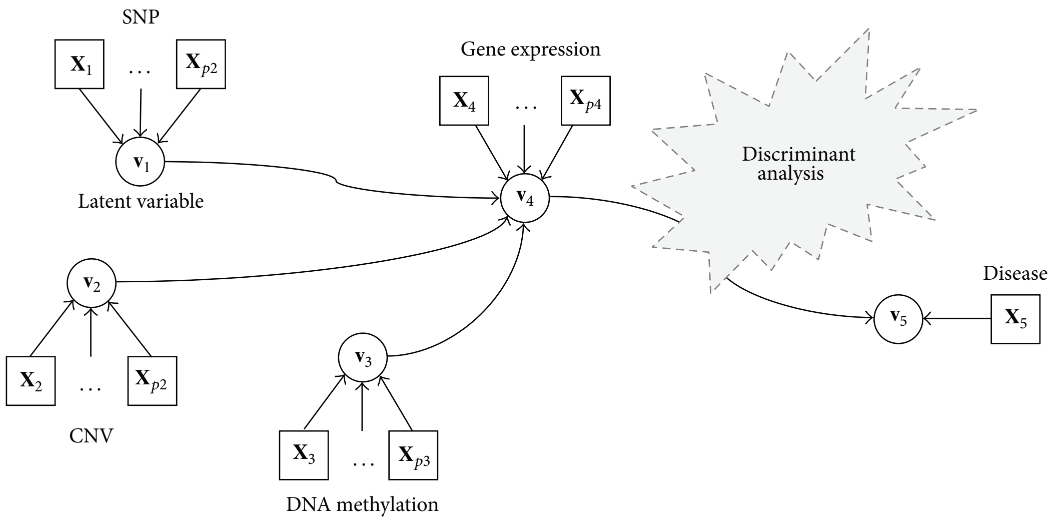

Figure 1: The conceptual graphic representation of the integrative genomic model. A rectangle represents a manipulated variable, and a circle represents a latent variable. The graphic representation illustrates the structure model that shows the relationship between SNP, CNV, DNA methylation, gene expression, and disease phenotype.

$\mathbf{X}_i$ and the ıth element of the vector $\mathbf{a}_i$, respectively. Figure 1 illustrates the conceptual overview of the multi-block data and framework.

2.2. Multiblock Discriminant Analysis. Multiblock Discriminant Analysis (MultiDA) builds a sparse association model by not only maximizing the total squared correlations between the multiblocks but also taking into account the discriminative factors in the model. MultiDA considers a linear subspace which is a construction of low-dimensional basis of the data. The linear subspaces of the multiblock, which maximize the total squared correlations, identify the significant factors of the association model with sparsity regularization. The linear subspace (or latent variable) $\mathbf{v}_j$ of the jth block is represented by

$$\mathbf{v}_j = \mathbf{X}_j \boldsymbol{\alpha}_j, \tag{1}$$

where $\boldsymbol{\alpha}_j$ is a loading vector. Then, we introduce sparse regularization (elastic net penalization) on the loading vector to reduce the chance of including insignificant variables and to improve their interpretation. The sparse regularization has its advantage especially when the number of features is much larger than the sample number ($N \ll P_j$). Therefore, the basic objective function can be represented as

$$\begin{aligned} &\underset{\boldsymbol{\alpha}_j}{\arg\max} \quad \sum_{j=1}^{J} \sum_{k=1, j\neq k}^{J} c_{jk} \frac{\boldsymbol{\alpha}_j^\top \mathbf{X}_j^\top \mathbf{X}_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_j^\top \mathbf{X}_j^\top \mathbf{X}_k \boldsymbol{\alpha}_k}{\boldsymbol{\alpha}_j^\top \mathbf{X}_j^\top \mathbf{X}_j \boldsymbol{\alpha}_j \boldsymbol{\alpha}_k^\top \mathbf{X}_k^\top \mathbf{X}_k \boldsymbol{\alpha}_k} \\ &\text{s.t.} \quad \boldsymbol{\alpha}_j^\top \mathbf{X}_j^\top \mathbf{X}_j \boldsymbol{\alpha}_j = 1, \\ &\qquad \left|\boldsymbol{\alpha}_j\right| \leq t_1, \\ &\qquad \left\|\boldsymbol{\alpha}_j\right\|^2 \leq t_2, \quad j = 1, \ldots, J, \end{aligned} \tag{2}$$

where $|\cdot|$ and $\|\cdot\|^2$ represent ℓ_1-norm and ℓ_2-norm of the vectors, respectively, and t_1 and t_2 are the shrinkage parameters that determine the sparsity. Note that the basic objective function is equivalent to the Sparse Generalized Canonical Correlation Analysis (SGCCA) [17]. Since the integrative genomic model aims to represent gene expression regulated by the combinations of SNP, CNV, and DM, the matrix $\mathbf{C}$ can be defined as

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3}$$

We further consolidate the model by (1) introducing a weight matrix of the correlation for the balance of the model and (2) providing discriminant analysis in the integrative genomic model. We also provide the sophisticated solution of the model while SGCCA heuristically estimates the optimal solution by following Wold's algorithm in the previous work [17].

2.2.1. Weight Matrix for the Balance of the Model. The weight matrix of the correlation between the multiblocks, $\mathbf{d} = \{d_{jk} \mid d_{jk} \in \Re, 1 \leq j, k \leq J\}$, is introduced in the model. In the original multiblock model, the correlation between gene expression and class label block tends to be overlooked. Instead, the sum of the squared pairwise correlations of $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, and $\mathbf{X}_4$ contributes large portions. The correlation weight matrix $\mathbf{D}$ gives an equal balance of the total squared correlations. In this paper, the weight matrix is defined as

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 3 \\ 0 & 0 & 0 & 3 & 0 \end{bmatrix}, \tag{4}$$

where the correlation between gene expression and class label blocks is three times more weighted than others. Then, the matrix $\mathbf{D}$ simply replaces the matrix $\mathbf{C}$.

2.2.2. Discriminant Analysis. In the proposed integrative genomic model, we need to find discriminative genes that characterize diseases. However, the integrative genomic model is comprised of combinations of multiple linear regression models. Thus, discriminant analysis such as Logistic Regression (LR) and Linear Discriminant Analysis (LDA) cannot be embedded into the integrative genomic model. To solve this problem, we adapted the Discriminative Least Squares Regression (DLSR) method proposed by Xiang et al. [19]. DLSR was developed based on the linear regression model, and it is proved that DLSR provides equal or superior performance compared to other discriminant methods. The basic concept of DLSR is to enlarge the distance between classes by introducing slack variables. Whereas they considered a multi-class problem and developed its sparse version with $\ell_{2,1}$-norm regularization in their work, we reformulated its sparse method with elastic net penalization to suit our own needs. In DLSR, the slack variable is introduced into the ordinary linear regression problem:

$$\mathbf{X}\mathbf{a} = \mathbf{y} + \mathbf{b} \odot \mathbf{m}, \tag{5}$$

where $\mathbf{y}$ is a dependent variable ($y_i = \{-1, 1\}, \mathbf{y} \in \mathfrak{R}^N$), $\mathbf{X}$ is a multivariate independent variable ($\mathbf{X} \in \mathfrak{R}^{N\times p}$), and $\mathbf{a}$ is a coefficient vector ($\mathbf{a} \in \mathfrak{R}^p$). $\mathbf{b}$ is a direction of the class, where its element $b_i = -1$ if $y_i = -1$ or 1 if otherwise ($\mathbf{b} \in \mathfrak{R}^p$). The Hadamard product operator $\odot$ of the direction vector $\mathbf{b}$ and the slack variable vector $\mathbf{m}$ determines the distance between classes ($\mathbf{m} \in \mathfrak{R}^p$). The optimal solution will be covered in the next section.

2.2.3. The Objective Function of MultiDA. We finally obtain the objective function of MultiDA:

$$\begin{aligned} \underset{\boldsymbol{\alpha}_j}{\arg\max} \quad & \sum_{j=1}^{J} \sum_{k=1,j\neq k}^{J} d_{jk} \frac{\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k}{\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j \boldsymbol{\alpha}_j \boldsymbol{\alpha}_k^\top \boldsymbol{\chi}_k^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k} \\ \text{s.t.} \quad & \boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j \boldsymbol{\alpha}_j = 1, \\ & \left|\boldsymbol{\alpha}_j\right| \leq t_1, \\ & \left\|\boldsymbol{\alpha}_j\right\|^2 \leq t_2, \quad j = 1, \dots, J, \end{aligned} \tag{6}$$

where $\boldsymbol{\chi}_j$ is defined as

$$\boldsymbol{\chi}_j = \begin{cases} \mathbf{X}_j + \mathbf{b} \odot \mathbf{m} & \text{if } j = 5 \\ \mathbf{X}_j & \text{if otherwise.} \end{cases} \tag{7}$$

This setting enables one to perform discriminant analysis between gene expression and disease blocks.

2.3. Optimization. The optimal solution of (6) can be obtained by the Lagrangian function:

$$\begin{aligned} \mathscr{L} = & -\sum_{j}^{J} \sum_{k=1,j\neq k}^{J} d_{jk} \boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k \boldsymbol{\alpha}_j{}^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k \\ & + \sum_{j}^{J} z_j \left(\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j \boldsymbol{\alpha}_j - 1 \right) + \sum_{j}^{J} \lambda_j \left|\boldsymbol{\alpha}_j\right| \\ & + \sum_{j}^{J} \frac{\left(1 - \lambda_j\right)}{2} \left\|\boldsymbol{\alpha}_j\right\|^2, \end{aligned} \tag{8}$$

where z_j and λ_j are the Lagrangian multipliers. The Lagrangian function (8) is convex, although not differentiable. Therefore, the local optimum of (8) provides a global solution. The partial derivatives of the Lagrangian function with respect to $\boldsymbol{\alpha}_j$ and λ_j are derived from

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial \boldsymbol{\alpha}_j} = & -\sum_{k}^{J} d_{jk} \left(\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k \right) \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k + z_j \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j \boldsymbol{\alpha}_j \\ & + \lambda_j \mathbf{s}_j + \left(1 - \lambda_j\right) \boldsymbol{\alpha}_j = 0, \end{aligned} \tag{9}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda_j} = \boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j \boldsymbol{\alpha}_j - 1 = 0, \tag{10}$$

where $\mathbf{s}_j$ is the vector of $\mathbf{a}_j$'s sign. Although the stationary equations have no closed form solutions, the optimal solution can be estimated by an iterative algorithm.

We can make (9) simple with the inner component:

$$\boldsymbol{v}_j = \sum_{k,k\neq j}^{J} d_{jk} \left(\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k \right) \boldsymbol{\chi}_k \boldsymbol{\alpha}_k. \tag{11}$$

Then, by introducing the inner component $\boldsymbol{v}_j$ into (9), the solution of $\boldsymbol{\alpha}_j$ can be written as

$$\boldsymbol{\alpha}_j = \left[z_j \left(\boldsymbol{\chi}_j^\top \boldsymbol{\chi}_j + \frac{1 - \lambda_j}{z_j} \right) \right]^{-1} \left(\boldsymbol{\chi}_j^\top \boldsymbol{v}_j - \lambda_j \mathbf{s}_j \right). \tag{12}$$

In (11), $(\boldsymbol{\alpha}_j^\top \boldsymbol{\chi}_j^\top \boldsymbol{\chi}_k \boldsymbol{\alpha}_k)$ is a squared correlation between the latent variables of the ith and jth block, which is a scalar. Therefore, the inner component is computed by $\boldsymbol{\alpha}_j$ of the previous iteration, and then new $\boldsymbol{\alpha}_j$ is updated in iterations.

Equation (12) is the normal equation of the regression of $\boldsymbol{v}_j$ on $\boldsymbol{\chi}_j$ with ridge and shrinkage parameter [20]. The final solution can be obtained by using the Univariate Soft-Thresholding (UST) method [21]:

$$\boldsymbol{\alpha}_{j_{(\imath)}} = \operatorname{sign}\left(\boldsymbol{\chi}_{j_{(\imath)}}^{\top}\boldsymbol{v}_j\right)\left(\left|\boldsymbol{\chi}_{j_{(\imath)}}^{\top}\boldsymbol{v}_j\right| - \lambda_j\right)_+, \tag{13}$$

where $\operatorname{sign}(x)$ returns a sign of x, that is, 1 if $x \geq 0$ or -1 if otherwise. $(x)_+$ returns only positive values of x (i.e., x if $x \geq 0$ or 0 if otherwise). λ_j can be obtained by K-fold cross-validation that minimizes mean squared errors. The parameter z_j can be ignored because the solution of $\boldsymbol{\alpha}_j$ is normalized by (10):

$$\boldsymbol{\alpha}_j = \frac{\sqrt{N}\boldsymbol{\alpha}_j}{\left\|\boldsymbol{\chi}_j\boldsymbol{\alpha}_j\right\|}. \tag{14}$$

For the discriminant analysis between gene expression and disease data blocks, the optimum of the slack variable $\mathbf{m}$ and the loading vector $\boldsymbol{\alpha}_4$ can be estimated by solving the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\alpha}_4,\mathbf{m}}{\arg\max} \quad & \frac{1}{2}\left\|\boldsymbol{\chi}_4\boldsymbol{\alpha}_4 - (\boldsymbol{v}_5 + \mathbf{b}\odot\mathbf{m})\right\|^2 \\ \text{s.t.} \quad & \left|\boldsymbol{\alpha}_4\right| \leq \xi_1, \\ & \left\|\boldsymbol{\alpha}_4\right\|^2 \leq \xi_2. \end{aligned} \tag{15}$$

The Lagrangian function of (15) is $\mathscr{L} = (1/2)\|\boldsymbol{\chi}_4\boldsymbol{\alpha}_4 - \boldsymbol{v}_5 - \mathbf{b}\odot\mathbf{m}\|^2 + \lambda_4|\boldsymbol{\alpha}_4| + ((1-\lambda_4)/2)\|\boldsymbol{\alpha}_4\|^2$. The derivative of the Lagrangian function with respect to $\boldsymbol{\alpha}_4$ is

$$\frac{\mathscr{L}}{\partial\boldsymbol{\alpha}_4} = \boldsymbol{\chi}_4^{\top}\boldsymbol{\chi}_4\boldsymbol{\alpha}_4 - \boldsymbol{\chi}_4^{\top}\boldsymbol{\gamma} + \lambda_4\mathbf{s} + (1-\lambda_4)\,\boldsymbol{\alpha}_4 = 0, \tag{16}$$

where $\mathbf{s}$ is the sign of $\boldsymbol{\alpha}_4$ and $\boldsymbol{\gamma} = \boldsymbol{v}_5 + \mathbf{b}\odot\mathbf{m}$. Thus, the equation of $\boldsymbol{\alpha}_4$ becomes

$$\boldsymbol{\alpha}_4 = \left(\boldsymbol{\chi}_4^{\top}\boldsymbol{\chi}_4 + 1 - \lambda_4\right)^{-1}\left(\boldsymbol{\chi}_4^{\top}(\boldsymbol{\gamma}) - \lambda_4\mathbf{s}\right). \tag{17}$$

Finally, the optimal solution of $\boldsymbol{\alpha}_4$ for the discriminative analysis is

$$\boldsymbol{\alpha}_{4_{(\imath)}} = \operatorname{sign}\left(\boldsymbol{\chi}_{4_{(\imath)}}^{\top}\boldsymbol{\gamma}\right)\left(\left|\boldsymbol{\chi}_{4_{(\imath)}}^{\top}\boldsymbol{\gamma}\right| - \lambda_4\right)_+. \tag{18}$$

λ_4 is also determined by K-fold cross-validation that minimizes mean squared errors like other λ_j's. The optimal solutions of $\mathbf{m}$ are simply derived from [19]

$$\mathbf{m} = \max\left(\mathbf{b}\odot\left(\boldsymbol{\chi}_4\boldsymbol{\alpha}_4 - \boldsymbol{v}_5\right), 0\right). \tag{19}$$

The brief algorithm is described in Algorithm 1. In the algorithm, r represents a rank of the subspace, which determines the dimension of the subspace. For instance, $\boldsymbol{\alpha}_j^r$ is rth rank of $\boldsymbol{\alpha}_j$. MultiDA optimizes the first rank subspace and iterates the optimization until the multiblock has no information. In lines 10–14 of Algorithm 1, Wold's procedure guarantees the convergence [22].

```
(1) For all block, normalize loading vectors
      $\boldsymbol{\alpha}_j^0 = \sqrt{N}\boldsymbol{\alpha}_j^0/|\boldsymbol{\chi}_j\boldsymbol{\alpha}_j|$
(2) $r = 1$
(3) **repeat**
(4)   **for** $j := 1$ **to** $J$ **do**
(5)     **for** $k := 1$ **to** $J$ **do**
(6)       **if** block $k$ is binary class data **then**
(7)         estimate $\mathbf{m}$ and $\boldsymbol{\alpha}_j$ by (18) and (19)
(8)         update $\boldsymbol{\chi}_k = \mathbf{X}_k + \mathbf{b}\odot\mathbf{m}$
(9)       **end if**
(10)      **if** $k < j$ **then**
(11)        $\boldsymbol{v}_j = \sum_{k=1,k\neq j}^{J} d_{jk}(\boldsymbol{\alpha}_j^{r\top}\boldsymbol{\chi}_j^{r\top}\boldsymbol{\chi}_k^{r}\boldsymbol{\alpha}_k^{r+1})\boldsymbol{\chi}_k^{r}\boldsymbol{\alpha}_k^{r+1}$
(12)      **else if** $k > j$ **then**
(13)        $\boldsymbol{v}_j = \sum_{k=1,k\neq j}^{J} d_{jk}(\boldsymbol{\alpha}_j^{r\top}\boldsymbol{\chi}_j^{r\top}\boldsymbol{\chi}_k^{r}\boldsymbol{\alpha}_k^{r})\boldsymbol{\chi}_k^{r}\boldsymbol{\alpha}_k^{r}$
(14)      **end if**
(15)      Compute $\boldsymbol{\alpha}_j^{r+1}$ by UST
            $\boldsymbol{\alpha}_{j_{(\imath)}}^{r+1} = \operatorname{sign}(\boldsymbol{\chi}_{j_{(\imath)}}^{\top}\boldsymbol{v}_j)(|\boldsymbol{\chi}_{j_{(\imath)}}^{\top}\boldsymbol{v}_j| - \lambda_j)_+$
(16)      Normalize $\boldsymbol{\alpha}_j^{r+1}$
            $\boldsymbol{\alpha}_j^{r+1} = \sqrt{n}\boldsymbol{\alpha}_j^{r+1}/|\boldsymbol{\chi}_j\boldsymbol{\alpha}_j^{r+1}|$
(17)      $r = r + 1$
(18)    **end for**
(19)  **end for**
(20) **until** $\sum_{j=1}^{J}\boldsymbol{\alpha}_j^r$ converges
```

Algorithm 1: Discriminant multiblock analysis.

3. Experiment Results

The goal of the assessment is to identify significant factors of the integrative genomic model with the multiblock data, specifically the discriminative factors of human disease. The discriminant factors include disease-specific locations or regions of SNP, CNV, DNA methylation, and gene expression against normal patients.

3.1. Simulation Study. We assessed the performance of the proposed method MultiDA through simulated data. Simulation data of various complexities were considered. Generation's schemes of the simulation data for the assessment were extended from the previous related works [16, 23].

Four generation functions of different complexity are defined as shown in Table 1. $\text{Type}_1(\mu)$ generates p-dimensional normally distributed random variables of a given mean (μ) and a variance $(\mathbf{I}_{p\times p})$, where $\mathbf{I}_{p\times p}$ is an $p \times p$ identity matrix. $\text{Type}_2(\mu,\delta)$ generates more complicated data than $\text{Type}_1(\mu)$. In $\text{Type}_2(\mu,\delta)$, a random model with a threshold (δ) is implemented with the function $\mathbf{1}_\delta$. Given a uniform distributed random value (u), $\mathbf{1}_\delta = 1$ if $u \leq \delta$ or 0 if otherwise. $\text{Type}_3(\mu,\rho)$ considers multicollinearity data in which more than two variables are highly correlated. The matrix data are generated by multivariate normal distribution $\mathcal{N}(\mu,\boldsymbol{\Sigma}_{p\times p})$. The covariance structure $\boldsymbol{\Sigma}_{p\times p}$ is built by the first order of autoregressive process. $\text{Type}_4(\mu,\sigma)$ generates p-dimensional normally distributed random variables from a given mean (μ) and a variance (σ).

Table 1: Generation functions.

Function	Model
$\text{Type}_1(\mu)$	$\mathbf{x} = \mu + \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})$
$\text{Type}_2(\mu, \delta)$	$\mathbf{x} = \mu + \mathbf{1}_\delta + \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})$
$\text{Type}_3(\mu, \rho)$	$\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma}_{p\times p})$
$\text{Type}_4(\mu, \sigma)$	$\mathbf{x} \sim \mathcal{N}(\mu, \sigma\mathbf{I}_{p\times p})$

Table 2: Scheme of the simulation data.

Simulation data	Generation model type	Column index
$\mathbf{X}_1$	$\mathbf{x}_i = \text{Type}_1(2.4)$	$1 \leq \imath \leq 5$
	$\mathbf{x}_i = \text{Type}_1(-2.6)$	$6 \leq \imath \leq 10$
	$\mathbf{x}_i = \text{Type}_2(1, 0.6)$	$11 \leq \imath \leq 40$
	$\mathbf{x}_i = \text{Type}_3(0, 0.8)$	$41 \leq \imath \leq 100$
$\mathbf{X}_2$	$\mathbf{x}_i = \text{Type}_1(3)$	$1 \leq \imath \leq 5$
	$\mathbf{x}_i = \text{Type}_1(4)$	$6 \leq \imath \leq 10$
	$\mathbf{x}_i = \text{Type}_3(0, 0.9)$	$11 \leq \imath \leq 60$
	$\mathbf{x}_i = \text{Type}_4(2, 2)$	$61 \leq \imath \leq 200$
$\mathbf{X}_3$	$\mathbf{x}_i = \text{Type}_1(5)$	$1 \leq \imath \leq 5$
	$\mathbf{x}_i = \text{Type}_1(-3)$	$6 \leq \imath \leq 10$
	$\mathbf{x}_i = \text{Type}_4(0, 1)$	$11 \leq \imath \leq 210$
	$\mathbf{x}_i = \text{Type}_3(0, 0.9)$	$211 \leq \imath \leq 300$

The first three multiblocks ($\mathbf{X}_j \in \mathfrak{R}^{N\times P_j}, 1 \leq j \leq 3$) were simulated by compounding the generation functions as defined in Table 2, where $P_1 = 100$, $P_2 = 200$, $P_3 = 300$, and $N = 500$. For instance, the first five columns of $\mathbf{X}_1$ were generated by $\text{Type}_1(2.4)$ and the following five columns were by $\text{Type}_1(-2.6)$. The next 30 columns were generated by the generation model with a threshold $\text{Type}_2(1, 0.6)$. The remaining columns of $\mathbf{X}_1$ were generated by the multicollinearity random variables $\text{Type}_3(0, 0.8)$. Then, we considered the multiblock linear model, $\mathbf{X}_4 = \sum_{j=1}^{3} \mathbf{X}_j\mathbf{B}_j + \Xi$, where $\mathbf{B}_j$ is a $P_j \times P_4$ loading matrix and Ξ is a $P_j \times P_4$ dimensional normally distributed noise matrix ($P_4 = 50$). We assumed that only the first ten variables of each block are significant to explain $\mathbf{X}_4$. The fifth block $\mathbf{X}_5$ is class label block. Given a coefficient vector $\mathbf{B}_4 \in \mathfrak{R}^{P_4\times 1}$ (all zeros but the first ten), the probability of disease π was computed by using

$$\pi = \frac{\exp\left(\mathbf{X}_4\mathbf{B}_4\right)}{1 + \exp\left(\mathbf{X}_4\mathbf{B}_4\right)}. \tag{20}$$

Then, the binary class label block was generated using the Bernoulli distribution with the probability π.

The simulation study was examined with 50 replications to assess the reproducibility. We compared the performance of MultiDA with the related methods, Sparse Canonical Correlation Analysis (SCCA) [24] and Sparse Generalized Canonical Correlation Analysis (SGCCA) [17]. SCCA is a two-block method that maximizes the correlation between independent $\mathcal{X}$ and response variable $\mathcal{Y}$. In SCCA, the three blocks of data were combined into a single block ($\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$), and the block GE was considered as response ($\mathcal{Y} = \mathbf{X}_4$). The class label block was not considered in SCCA. The multiblock method SGCCA was tuned to be compatible with the proposed integrative genomic model. Note that the same matrix $\mathbf{C}$ was used in SGCCA, but SGCCA did not take the discriminant analysis into account.

We examined the performance by how well they correctly identify significant factors of the integrative association model. Given a ground truth, we computed a confusion matrix and measured True Positive Rate (TPR), Positive Predictive Value (PPV), and Accuracy (ACCU). In the sparse setting, the true negatives are relatively much larger than false positives. Therefore, True Negative Rates (TNR) and Negative Predictive Values (NPV) were not included in this paper. The results of the simulation experiment are illustrated in Figure 2. The proposed method MultiDA (0.93 ± 0.03) and the multiblock method SGCCA (0.93 ± 0.03) outperformed SCCA (0.83 ± 0.24) in terms of TPR. It supports that the multiblock methods reduce false negatives that incorrectly identify the significant as the insignificant. MultiDA appeared as the best performance in PPV and ACCU. MultiDA produced 0.58 ± 0.07 and 0.95 ± 0.01 for PPV and ACCU, respectively. Higher PPV values represent lower false positives that incorrectly identify the insignificant as the significant. The PPV and ACCU of SCCA were 0.48 ± 0.15 and 0.89 ± 0.14 and were 0.54 ± 0.08 and 0.94 ± 0.01 for SGCCA, respectively.

3.2. Human Brain Data of Schizophrenia.

Human brain data were obtained from three major psychiatric disorders such as schizophrenia (SZ), bipolar disorder (BP), and major depression (DP) as well as from control group. Specifically, 39 samples of SZ, 35 samples of BP, 12 samples of DP, and 43 samples of control were provided from the Stanley Medical Research Institute. SNP, CNV, DNA methylation, and gene expression data were acquired from the human prefrontal cortex of the 129 samples in the preparation of this experiment. For each individual, 10,760 SNPs after removing highly correlated ones, 1,028 CNVs, 20,769 DNA methylations, and 19,767 gene expressions were examined. Due to the recent research that reported that genetic effects may be largely shared in major psychiatric disorders such as autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia, we considered those psychiatric diseases together and performed MultiDA to identify discriminate factors against the control [25, 26].

The multiblock data was analyzed by MultiDA. As a result of the analysis, 78 SNPs, 30 CNVs, 47 DNA methylations, and 35 genes were detected, where the high correlation between the connections was found. The potential gene markers of the psychiatric disorders were inferred from the result of the proposed method. The genes physically located near the selected SNPs and the genes corresponding to the result of CNV and the DNA methylation were chosen. Significantly observed genes among the results of MultiDA are listed in Table 3, where the data source of the gene and literature regarding the psychiatric disorders are described.

The gene regulatory network of the genes from the result was searched by STRING database [27]. Among a number of the retrieved interactions, we take note of one gene

Table 3: The gene results from MultiDA with psychiatric disorders.

Gene	Chromosome	Location	Source	ID	MAF	Reference
HTR7	10	10q21-q24	GE	7934970		[28]
APOE	19	19q13.2	DM	cg14123992		[29]
TRPM1	15	15q13.3	DM	cg18085517		
EPHB1	3	3q21-q23	CNV	CNP12652		
NPY	7	7p15.1	CNV	CNP2267		[30]
QKI	6	6q26	SNP	rs1336225	0.18	
SLC15A1	13	13q32.3	SNP	rs9517421	0.17	[31]
NPAS3	14	14q13.1	SNP	rs1124910	0.25	[32]
C15orf53	15	15q14	SNP	rs1433876	0.29	[33]

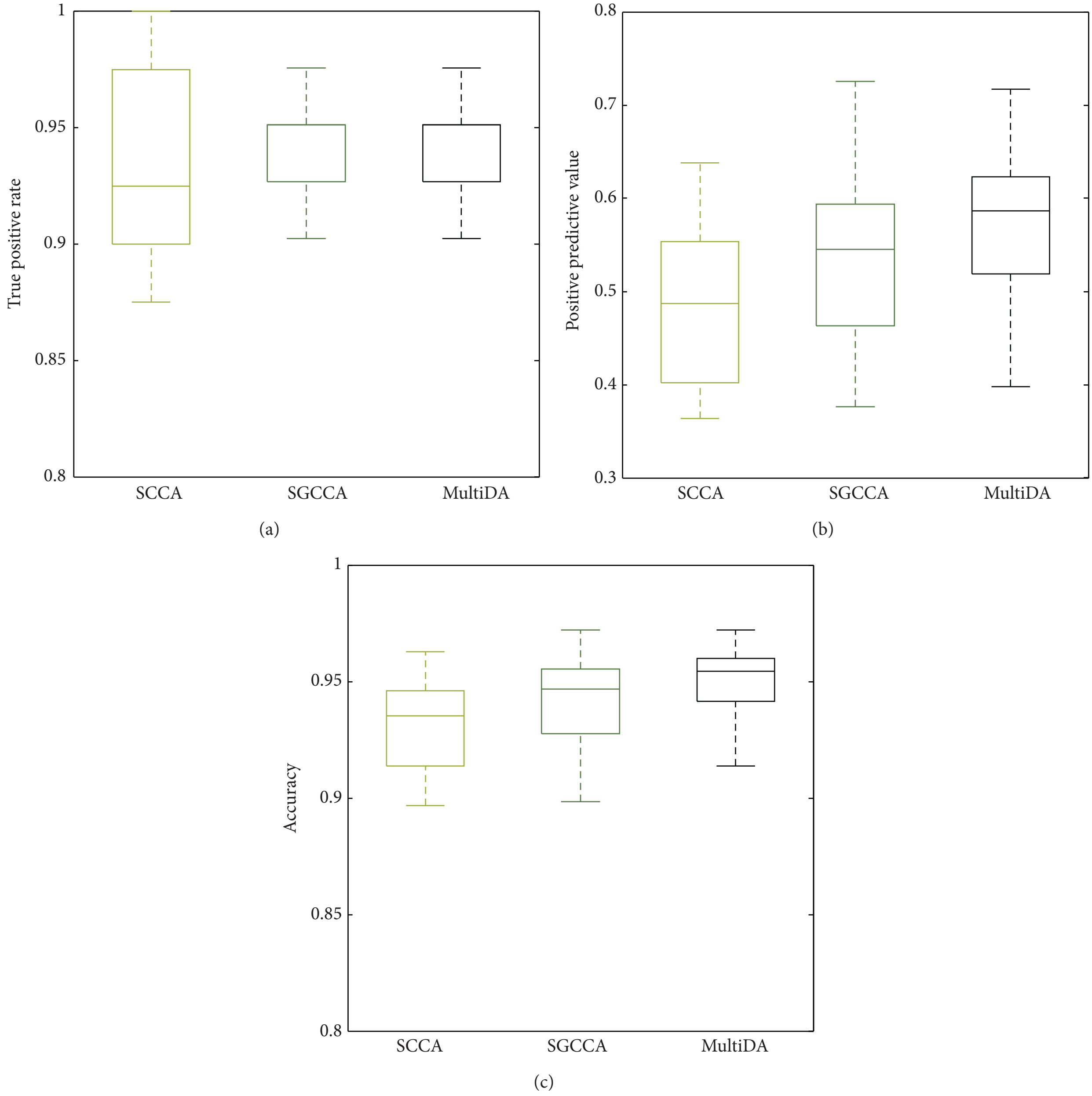

Figure 2: Performance comparison in simulation study: (a) True Positive Rate; (b) Positive Predictive Value; (c) Accuracy.

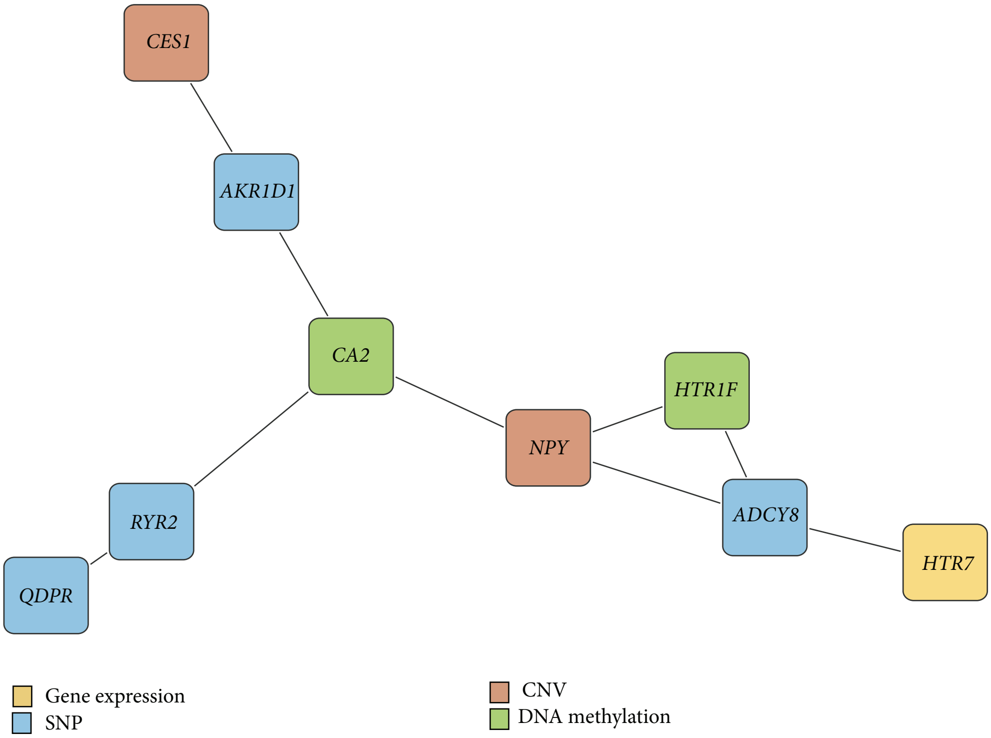

Figure 3: The gene regulatory network searched with the gene results by STRING database. The legend shows the data source of the gene.

regulatory network illustrated in Figure 3. The interaction network consists of *HTR7*, *ADCY8*, *HTR1F*, *NPY*, *CA2*, *RYR2*, *QDPR*, *AKR1D1*, and *CES1* gene. *HTR7* is inferred from the gene expression set, *HTR1F* and *CA2* are from the DNA methylation expression, *NPY* and *CES1* are from the CNV, and the others are from the SNP data. The negative coefficient of *HTR1F* in the model may support the widely accepted notion that DNA methylation suppresses gene regulation impeding the binding of transcriptional proteins to the gene [34]. In particular, the *HTR7* gene (5-hydroxytryptamine receptor 7) is a major neurotransmitter in the central nervous system, and a number of literatures related to bipolar and schizophrenia disorder are reported [28]. Interestingly, the *HTR7* gene was found in the gene expression data block in this study, while the other previous researches reported the gene with GWAS on the SNP data block. The gene may have strong incorporated interactions with other heterogeneous data, which is consequently considered to be significant in the integrative model. It supports the strength of the integrative approach. Moreover, we found that *HTR7* and *NPY* are in the same pathway, which is *neuroactive ligand-receptor interaction*, where the *NPY* gene is also a neurotransmitter in the brain and is known to play an important role in the emotional process [30]. A large number of psychiatric disorder susceptible genes were associated with this pathway [25]. *ADCY8*, which interacts with both *HTR7* and *NPY*, may be potentially a susceptibility gene that causes the psychiatric disorders. In previous research [35], they found that *ADCY8* is a susceptibility gene for avoidance behavior on mouse and also found that it indirectly induces the susceptibility on human mood disorders. Our result supports their claim.

4. Conclusion

In this paper, we developed the novel Multiblock Discriminant Analysis method in order to dissect the mechanism of complex human disease using multiple genetic data. The genomic association study with single type data may fall short of identifying the mechanisms of the diseases. On the other hand, MultiDA enables comprehensive analysis using multiple genetic data. Moreover, MultiDA provides analysis for the special setting of binary class data, where it greatly detects discriminative factors in the integrative genomic model. The simulation experiments support the outstanding performance of the proposed methods. As a target application, psychiatric disorder disease data, including SNP, CNV, DNA methylation, and gene expression, were analyzed in the integrative genomic model. Among the large number of variables of each block, candidate biomarkers were proposed as significant components of the disease mechanism. The proposed methods capture the global profile of the mechanism that conventional single or two block methods fail to detect. This promising tool for the integrative genomic study can provide flexible extensibility for new types of data in the era, superseding new high-throughput technologies.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[2] C. N. Henrichsen, E. Chaignat, and A. Reymond, "Copy number variants, diseases and gene expression," *Human Molecular Genetics*, vol. 18, no. 1, pp. R1–R8, 2009.

[3] Y. Gilad, S. A. Rifkin, and J. K. Pritchard, "Revealing the architecture of gene regulation: the promise of eQTL studies," *Trends in Genetics*, vol. 24, no. 8, pp. 408–415, 2008.

[4] M. Slatkin, "Epigenetic inheritance and the missing heritability problem," *Genetics*, vol. 182, no. 3, pp. 845–850, 2009.

[5] J. L. Freeman, G. H. Perry, L. Feuk et al., "Copy number variation: new insights in genome diversity," *Genome Research*, vol. 16, no. 8, pp. 949–961, 2006.

[6] S. Girirajan, C. D. Campbell, and E. E. Eichler, "Human copy number variation and complex genetic disease," *Annual Review of Genetics*, vol. 45, pp. 203–226, 2011.

[7] E. N. Gal-Yam, Y. Saito, G. Egger, and P. A. Jones, "Cancer epigenetics: modifications, screening, and therapy," *Annual Review of Medicine*, vol. 59, pp. 267–280, 2008.

[8] L. D. Moore, T. Le, and G. Fan, "DNA methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2013.

[9] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[10] M. R. Aure, S.-K. Leivonen, T. Fleischer et al., "Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors," *Genome Biology*, vol. 14, no. 11, article R126, 2013.

[11] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, vol. 15, no. 2, article R37, 2014.

[12] A. C. Nica, S. B. Montgomery, A. S. Dimas et al., "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations," *PLoS Genetics*, vol. 6, no. 4, Article ID e1000895, 2010.

[13] Y.-H. Hsu, M. C. Zillikens, S. G. Wilson et al., "An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits," *PLoS Genetics*, vol. 6, no. 6, Article ID e1000977, 2010.

[14] Q. Xiong, N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, "Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets," *Genome Research*, vol. 22, no. 2, pp. 386–397, 2012.

[15] L. Conde, P. M. Bracci, R. Richardson, S. B. Montgomery, and C. F. Skibola, "Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma," *American Journal of Human Genetics*, vol. 92, no. 1, pp. 126–130, 2013.

[16] W. Li, S. Zhang, C. C. Liu, and X. J. Zhou, "Identifying multi-layer gene regulatory modules from multi-dimensional genomic data," *Bioinformatics*, vol. 28, no. 19, Article ID bts476, pp. 2458–2466, 2012.

[17] M. Kang, B. Zhang, X. Wu, C. Liu, and J. Gao, "Sparse generalized canonical correlation analysis for biological model integration: a genetic study of psychiatric disorders," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, pp. 1490–1493, July 2013.

[18] Q. Zhao, X. Shi, Y. Xie, J. Huang, B. Shia, and S. Ma, "Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA," *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 291–303, 2015.

[19] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1738–1754, 2012.

[20] A. Tenenhaus and M. Tenenhaus, "Regularized generalized canonical correlation analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, 2011.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B:ethodological*, vol. 58, no. 1, pp. 267–288, 1996.

[22] M. Hanafi, "PLS path modelling: computation of latent variables with the estimation mode B," *Computational Statistics*, vol. 22, no. 2, pp. 275–292, 2007.

[23] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, 2008.

[24] S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman, "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, 2008.

[25] P. Ragunath, R. Chitra, S. Mohammad, and P. Abhinand, "A systems biological study on the comorbidity of autism spectrum disorders and bipolar disorder," *Bioinformation*, vol. 7, no. 3, pp. 102–106, 2011.

[26] A. Serretti and C. Fabbri, "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *The Lancet*, vol. 381, no. 9875, pp. 1371–1379, 2013.

[27] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[28] Y. M. J. Lin, H. C. Yang, T. J. Lai, C. S. J. Fann, and H. S. Sun, "Receptor mediated effect of serotonergic transmission in patients with bipolar affective disorder," *Journal of Medical Genetics*, vol. 40, no. 10, pp. 781–786, 2003.

[29] F. Vila-Rodriguez, W. G. Honer, S. M. Innis, C. L. Wellington, and C. L. Beasley, "ApoE and cholesterol in schizophrenia and bipolar disorder: comparison of grey and white matter and relation with APOE genotype," *Journal of Psychiatry & Neuroscience*, vol. 36, no. 1, pp. 47–55, 2011.

[30] M. Heilig, "The NPY system in stress, anxiety and depression," *Neuropeptides*, vol. 38, no. 4, pp. 213–224, 2004.

[31] M. Maheshwari, S. L. Christian, C. Liu et al., "Mutation screening of two candidate genes from 13q32 in families affected with bipolar disorder: human peptide transporter (SLC15A1)

and human glypican5 (GPC5)," *BMC Genomics*, vol. 3, article 30, 2002.

[32] B. S. Pickard, A. Christoforou, P. A. Thomson et al., "Interacting haplotypes at the NPAS3 locus alter risk of schizophrenia and bipolar disorder," *Molecular Psychiatry*, vol. 14, no. 9, pp. 874–884, 2009.

[33] T. M. Kranz, S. Ekawardhani, M. K. Lin et al., "The chromosome 15q14 locus for bipolar disorder and schizophrenia: is *C15orf53* a major candidate gene?" *Journal of Psychiatric Research*, vol. 46, no. 11, pp. 1414–1420, 2012.

[34] P. A. Jones, "Functions of DNA methylation: islands, start sites, gene bodies and beyond," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 484–492, 2012.

[35] A. G. de Mooij-van Malsen, H. A. van Lith, H. Oppelaar et al., "Interspecies trait genetics reveals association of Adcy8 with mouse avoidance behavior and a human mood disorder," *Biological Psychiatry*, vol. 66, no. 12, pp. 1123–1130, 2009.