A Class-Information-Based Sparse Component Analysis Method to Identify Differentially Expressed Genes on RNA-Seq Data

Jin-Xing Liu,
Yong Xu, Ying-Lian Gao,
Chun-Hou Zheng,
Dong Wang, and Qi Zhu

Abstract—With the development of deep sequencing technologies, many RNA-Seq data have been generated. Researchers have proposed many methods based on the sparse theory to identify the differentially expressed genes from these data. In order to improve the performance of sparse principal component analysis, in this paper, we propose a novel class-information-based sparse component analysis (CISCA) method which introduces the class information via a total scatter matrix. First, CISCA normalizes the RNA-Seq data by using a Poisson model to obtain their differential sections. Second, the total scatter matrix is gotten by combining the between-class and within-class scatter matrices. Third, we decompose the total scatter matrix by using singular value decomposition and construct a new data matrix by using singular values and left singular vectors. Then, aiming at obtaining sparse components, CISCA decomposes the constructed data matrix by solving an optimization problem with sparse constraints on loading vectors. Finally, the differentially expressed genes are identified by using the sparse loading vectors. The results on simulation and real RNA-Seq data demonstrate that our method is effective and suitable for analyzing these data.

Index Terms—Constrained optimization, feature selection, multivariate statistics, principal component analysis, singular value decomposition

1 Introduction

It is one of the challenges in current molecular biology to find the genes associated with specific biological functions or cellular processes. Up to date, these genes have been detected more comprehensively than ever before by deep sequencing (also called next-generation sequencing) technologies [1]. These technologies have generated many data which make it possible to monitor gene expression levels on a genomic scale and to understand the mechanism of life [2], [3]. RNA-Seq uses deep sequencing technologies to sequence cDNA that has been derived from a RNA experiment, and hence produces millions of short reads. These reads are then typically mapped to a reference genome and the number of reads mapping within a genomic feature of interest (such as a gene or an exon) is used to quantify gene expression in the analyzed sample [4].

- *J. -X. Liu is with the College of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 276826, China, and the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China. E-mail: sdcavell@126.com.*
- *Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, and the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, Guangdong 518055, China. E-mail: yongxu@ymail.com.*
- *Y. -L. Gao is with the Library of Qufu Normal University, Qufu Normal University, Rizhao 276826, China. E-mail: yinliangao@126.com.*
- *C. -H. Zheng is with the College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230039, China. E-mail: zhengch99@126.com.*
- *D. Wang is with the College of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 276826, China. E-mail: dongwshark@126.com.*
- *Q. Zhu is with the Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangshu 210016, China. E-mail: zhuqi@nuaa.edu.cn.*

Manuscript received 15 Mar. 2014; revised 16 Oct. 2014; accepted 20 May 2015. Date of publication 17 June 2015; date of current version 30 Mar. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2440265

Because the nature of RNA-Seq data can result in different experiments with dramatically different total number of sequence reads, counts from each experiment should be normalized by the sequencing depth of the corresponding experiment before any comparison is made between experiments [5-8].

Generally speaking, only the differentially expressed (DE) genes are fundamentally concerned by biologists. Although expression profiles of thousands of genes are simultaneously measured, most genes expression profiles are flat and only a small number of gene expression profiles are differential. Therefore, it is a matter of great urgency to effectively identify the DE genes from these RNA-Seq data.

Feature selection is the simplest method to identify the DE genes. It first calculates a score for each feature, and then selects the features with high scores [9, 10]. For example, PoissonSeq (PS) method was proposed to identify the DE genes by Li et al. [7]. Because feature selection method separately calculates a score of each feature, it ignores dependencies among features. The methods of feature extraction have been proposed [11-14] to overcome this shortcoming. Unlike feature selection, feature extraction simultaneously uses all the data information. Among feature extraction methods, principal component analysis (PCA) [11] is the most commonly used one. Classical PCA is defined as follows:

$$z^* = \arg\max_{z^T z \le 1} \|\mathbf{F}z\|_2, \tag{1}$$

where $\mathbf{F} \in \mathrm{R}^{n \times m}$ is a data matrix encoding n samples of m variables.

PCA has been used to analyze the deep sequencing data. For example, Ji et al. used PCA for analyzing multiple deep sequencing datasets to identify the differential protein-DNA interactions between two biological conditions [15]. Pickrell et al. used PCA to quantile-normalize the RNA-Seq data and to remove the confounding effects [16]. Singh et al. used PCA to investigate the relationships between the chickpea tissues on the whole-gene expression data set [17].

A good analysis tool for biological interpretation should be able to highlight 'simple' structures in the genome-structures. The 'simple' structures are expected to involve only a few genes that are associated with a specific biological function or process [18]. The objective of sparse PCA (SPCA) is to make a trade-off between statistical fidelity and interpretability. In recent years, many SPCA methods have been proposed to maximize the explained variance. For example, Journée et al. proposed a SPCA method by using generalized power method [18]. Zou et al. viewed SPCA as a regression optimization problem and imposed the LASSO penalty on the regression coefficients [19]. Shen et al. used the singular value decomposition (SVD) to obtain a low-rank matrix approximation of a data matrix via sparse penalties [20]. Witten et al. proposed a penalized matrix decomposition (PMD) via sparse penalties [21].

As mentioned above, only a small number of genes are differentially expressed in gene expression data, so these sparse methods can meet with requirements for analyzing these data. For example, Lee et al. used SPCA to analyze high-throughput genomic data [22]. In the case of DNA methylation, Zhuang et al. showed SPCA outperforms many other feature selection methods [23].

Although these sparse methods have been widely used for analyzing gene expression data, they have some deficiencies. For instance, when class labels of samples have been known, these methods cannot take advantage of the class information, because they are unsupervised. In this paper, we propose a method of class-information-based sparse component analysis (CISCA) to improve the analytical performance for RNA-Seq data. CISCA introduces the class information of samples by using a total scatter matrix. The scheme of CISCA is given as follows: First of all, CISCA normalizes RNA-Seq data to obtain the

differential section by using a Poisson model. Second, we get a total scatter matrix based on the differential section of RNA-Seq data. Third, CISCA decomposes the total scatter matrix by using singular value decomposition and constructs a new data matrix by using singular values and left singular vectors. Fourth, CISCA decomposes the constructed data matrix to obtain the sparse components by solving an optimization problem with sparse constraints on loading vectors. Finally, the DE genes are identified by using the sparse loading vectors.

The main contributions of our work are as follows: first, it proposes, for the first time, the idea and method of CISCA for analyzing RNA-Seq data; second, it introduces θ to the constructed matrix $\mathbf{F}$, which strengthens strong signals and weakens weak signals; third, it brings class labels of samples into sparse PCA by total scatter matrix; finally, it provides a large number of experiments on simulation and real RNA-Seq data sets.

The rest of this paper is organized as follows. The methodology of CISCA is shown in Section 2. Section 3 gives the results and discussion. Section 4 concludes this paper.

2 Methodology

In this section, the method of class-information-based sparse component analysis is proposed.

2.1 Normalizing RNA-Seq Data

Assuming that all the data points are stacked as column vectors of a matrix $\mathbf{X}$ with size $m \times n$, in general, $m \gg n$. In the case of RNA-Seq count data, for the sample j, x_{ij} is the number of reads overlapping gene i included in the Ensembl annotation of the given organism's genome [24]. Let $\mathbf{Y}$ denote a class labels vector. Assuming that $x_{ij} \tilde{\text{Poisson}}(\xi_{ij})$, where the form of ξ_{ij} is given by a log-linear model as follows [7]:

$$\log \xi_{ij} = \log d_j + \log \beta_i + e_{ij}, \tag{2}$$

where

d_j: the sequencing depth of sample j;
β_i: the non-differential expression of gene i;
e_{ij}: the differential expression of gene i in sample j.

Without loss of generality, we assume that $\sum_{j=1}^{n} d_j = 1$. And β_i can be calculated by $\beta_i = \sum_{j=1}^{n} x_{ij}$.

First, CISCA estimates the sequencing depth d_j by using the Li's method in [7]. Then, the differential expression e_{ij} of gene i in sample j can be calculated as follows:

$$e_{ij} = \log \xi_{ij} - \log d_j - \log \beta_i. \tag{3}$$

After obtaining the element e_{ij} of differential expression matrix $\mathbf{E}$, we analyze the matrix $\mathbf{E}$ to identify DE genes by using feature extraction method.

2.2 Definition of Scatter Matrices

Mathematically speaking, for all the samples of all classes, three measures are defined: (a) the first one is between-class scatter matrix which is given by

$$\mathbf{S}_b = \sum_{j=1}^{c} N_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T, \tag{4}$$

where μ_j is the mean of class j, $\bar{\mu}$ represents the mean of all classes, c is the number of classes, N_j is the number of samples in class j; (b) the second one is within-class scatter matrix which is given as follows:

$$\mathbf{S}_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{e}_i^j - \mu_j)(\mathbf{e}_i^j - \mu_j)^T, \tag{5}$$

where $\mathbf{e}_i^j$ is the sample i of class j; (c) the total scatter matrix can be defined as follows [25]:

$$\mathbf{S}_t = \mathbf{S}_b - \eta \mathbf{S}_w, \tag{6}$$

where $\eta \geq 0$ is a regulation parameter which gives a trade-off between $\mathbf{S}_b$ and $\mathbf{S}_w$.

Let $\mathbf{G} = \mathbf{W}\Sigma\mathbf{H}^T$ be the SVD of $\mathbf{G}$, where $\mathbf{G}$ is defined as $\mathbf{S}_t = \mathbf{G}\mathbf{G}^T$, $\mathbf{W}$ and $\mathbf{H}$ are orthogonal, $\Sigma = \begin{pmatrix} \Sigma_t & 0 \\ 0 & 0 \end{pmatrix}$, $\Sigma_t \in R^{t \times t}$ is diagonal, and $t = \text{rank}(\mathbf{S}_t)$. Then

$$\begin{aligned} \mathbf{S}_t &= \mathbf{G}\mathbf{G}^T \\ &= \mathbf{W}\Sigma\mathbf{H}^T\mathbf{H}\Sigma^T\mathbf{W}^T \\ &= \mathbf{W}\Sigma\Sigma^T\mathbf{W}^T \\ &= \mathbf{W}\begin{pmatrix} \Sigma_t & 0 \\ 0 & 0 \end{pmatrix}\mathbf{W}^T. \end{aligned} \tag{7}$$

Let $\mathbf{W} = (\mathbf{W}_1 \quad \mathbf{W}_2)$ be a partition of $\mathbf{W}$, such that $\mathbf{W}_1 \in R^{m \times t}$ and $\mathbf{W}_2 \in R^{m \times (m-t)}$. Since $\mathbf{S}_t = \mathbf{S}_b - \eta \mathbf{S}_w$, we have

$$\begin{aligned} \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} &= \mathbf{W}^T(\mathbf{S}_b - \eta \mathbf{S}_w)\mathbf{W} \\ &= \begin{pmatrix} \mathbf{W}_1^T \\ \mathbf{W}_2^T \end{pmatrix} \mathbf{S}_b (\mathbf{W}_1 \quad \mathbf{W}_2) - \eta \begin{pmatrix} \mathbf{W}_1^T \\ \mathbf{W}_2^T \end{pmatrix} \mathbf{S}_w (\mathbf{W}_1 \quad \mathbf{W}_2) \\ &= \begin{pmatrix} \mathbf{W}_1^T\mathbf{S}_b\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{S}_b\mathbf{W}_2 \\ \mathbf{W}_2^T\mathbf{S}_b\mathbf{W}_1 & \mathbf{W}_2^T\mathbf{S}_b\mathbf{W}_2 \end{pmatrix} - \eta \begin{pmatrix} \mathbf{W}_1^T\mathbf{S}_w\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{S}_w\mathbf{W}_2 \\ \mathbf{W}_2^T\mathbf{S}_w\mathbf{W}_1 & \mathbf{W}_2^T\mathbf{S}_w\mathbf{W}_2 \end{pmatrix}. \end{aligned} \tag{8}$$

It follows that $\mathbf{W}_2^T\mathbf{S}_b\mathbf{W}_2 - \eta\mathbf{W}_2^T\mathbf{S}_w\mathbf{W}_2 = 0$. Therefore

$$\eta = trace(\mathbf{W}_2^T\mathbf{S}_b\mathbf{W}_2)/trace(\mathbf{W}_2^T\mathbf{S}_w\mathbf{W}_2), \tag{9}$$

where the traces can measure the distances of between-class and within-class scatter matrices [26].

We thus have

$$\begin{cases} \mathbf{W}_2^T\mathbf{S}_b\mathbf{W}_1 - \eta\mathbf{W}_2^T\mathbf{S}_w\mathbf{W}_1 = 0, \\ \mathbf{W}_1^T\mathbf{S}_b\mathbf{W}_2 - \eta\mathbf{W}_1^T\mathbf{S}_w\mathbf{W}_2 = 0. \end{cases} \tag{10}$$

Here, similarly with PCA, the off-diagonal and diagonal elements of $\mathbf{S}_t$ may reflect the covariance between any two samples and variance of the samples, respectively.

2.3 The Definition of Optimization Problem

To help improve the analytical performance for RNA-Seq data, CISCA introduces the class information via the total scatter matrix $\mathbf{S}_t$.

First, the total scatter matrix $\mathbf{S}_t$ is decomposed by SVD, as given by

$$\mathbf{S}_t = \mathbf{U}\Delta\mathbf{V}^T, \tag{11}$$

where $\mathbf{U}$, $\mathbf{V}$ are orthogonal and $\Delta = diag(\delta_1, \delta_2, \cdots, \delta_r)$ is a diagonal matrix of the singular values, r is the rank of $\mathbf{S}_t$.

Then, CISCA constructs a new data matrix as follows:

$$\mathbf{F} = (\mathbf{U}\Delta^{\theta})^T, \tag{12}$$

where $\theta \geq 0$ is a scaling parameter. When $\theta = 0$, $\mathbf{F} = \mathbf{U}^T$. Because the matrices $\mathbf{S}_b$ defined in (4) and $\mathbf{S}_w$ defined in (5) are symmetic, which will lead to a symmetic 'total scatter matrix' $\mathbf{S}_t$, the left and right eigenvector matrices in (11) should be the same, implying that the transformed $\mathbf{F}$ matrix should have the same eigenvectors with $\mathbf{S}_t$. This also means that when $\theta = 0.5$, performing sparse PCA on $\mathbf{F}$ would be equivalent to running sparse PCA on $\mathbf{S}_t$.

We can find a reasonable sparse loading vector z by solving an optimization problem under the sparse l_1-norm constraint. The optimization problem can be written as the following formulation:

$$z = \arg\max_{z \in B^m} \|\mathbf{F}z\|_2 - \gamma\|z\|_1, \tag{13}$$

where $\gamma \geq 0$ is a sparsity-controlling parameter and $B^m = \{z \in \mathbf{R}^m | z^T z \leq 1\}$ is referred to the unit Euclidean ball in $\mathbf{R}^m$.

2.4 The Solution to Optimization Problem

When $\gamma = 0$, the optimization problem (13) degrades to the classical problem (1). The solution of problem (13) is equal to the first principal component of the data matrix $\mathbf{F}$.

For any $z \neq 0$, the following inequality can be deduced [18]:

$$\begin{aligned} \max\|\mathbf{F}z\|_2/\|z\|_1 &= \max\Big\|\sum\nolimits_i z_i f_i\Big\|_2 \Big/ \Big(\sum\nolimits_i |z_i|\Big) \\ &\leq \max\Big(\sum\nolimits_i |z_i| \|f_i\|_2\Big) \Big/ \Big(\sum\nolimits_i |z_i|\Big) \\ &= \max_i \|f_i\|_2 = \|f_i\|_2^{max}. \end{aligned} \tag{14}$$

Provided that $\gamma \geq \|f_i\|_2^{max}$, we have $\|\mathbf{F}z\|_2 - \gamma\|z\|_1 \leq 0$ for any $z \neq 0$. So the range of sparsity-controlling parameter γ should be $\big[0, \|f_i\|_2^{max}\big)$.

In the case when $r << m$, the feasible set of problem (13) is of high dimension. According to Journée et al. [18], the following equation can be deduced:

$$\begin{aligned} z &= \arg\max_{z \in B^m} \max_{x \in B^r} x^T \mathbf{F} z - \gamma\|z\|_1 \\ &= \arg\max_{x \in B^r} \max_{z \in B^m} \sum\nolimits_{i=1}^m \big[z_i \big(f_i^T x\big) - \gamma |z_i|\big]. \end{aligned} \tag{15}$$

Let $\hat{z}_i = \operatorname{sign}\big(f_i^T x\big) z_i$ (the sign function can obtain the sign of the argument), the optimization problem (15) can be reformulated as follows:

$$\hat{z} = \arg\max_{x \in B^r} \max_{z \in B^m} \sum\nolimits_{i=1}^m |\hat{z}_i| \big(\big|f_i^T x\big| - \gamma\big). \tag{16}$$

When $\gamma \in \big[0, \|f_i\|_2^{max}\big)$, there is some $x \in B^m$. Fixing x, the closed-form solution of the problem (16) can be gotten for $\hat{z}$:

$$\hat{z}_i = \big\lfloor |f_i^T x| - \gamma \big\rfloor \Big/ \sqrt{\sum\nolimits_{k=1}^m \big\lfloor |f_i^T x| - \gamma \big\rfloor^2}, i = 1, \cdots, m, \tag{17}$$

where $\lfloor g \rfloor = \max\{0, g\}$. Then transforming $\hat{z}$ into z, the following solution can be obtained [18]:

$$z_i = \operatorname{sgn}\big(f_i^T x\big) \big\lfloor |f_i^T x| - \gamma \big\rfloor \Big/ \sqrt{\sum\nolimits_{k=1}^m \big\lfloor |f_i^T x| - \gamma \big\rfloor^2}, i = 1, \cdots, m \tag{18}$$

Journée et al. proposed a generalized power method to solve the optimization problem [18].

2.5 Multiple Components

CISCA can obtain multiple components by using deflation techniques which solve the optimization problem (13) repeatedly, and each time use the residual which can be obtained by subtracting the part found previously from the original matrix. According to d'Aspremont [27], the residual can be obtained as follows:

$$res = \mathbf{F} - qz^T, \tag{19}$$

where $q = \mathbf{F}z$ is the vector that solves

$$\min\big\|\mathbf{F} - qz^T\big\|_F. \tag{20}$$

Further deflation techniques for PCA have been proposed by Mackey [28].

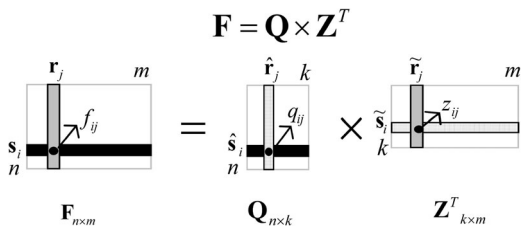

Fig. 1. The graphical depiction of CISCA of the matrix **F**, with factor scores **Q** and PCs **Z**. $\tilde{\mathbf{r}}_j$ is the row vector of PCs **Z**, which transforms the data vector $\mathbf{r}_j$ into factor scores $\hat{\mathbf{r}}_j$. Correspondingly, $\tilde{\mathbf{s}}_j$ is the column vector of PCs **Z**, which transforms the data vector $\mathbf{s}_j$ into factor scores $\hat{\mathbf{s}}_j$.

2.6 The Algorithm

The algorithm of CISCA is shown as follows.

Algorithm 1. CISCA Algorithm

Input: Data matrix $\mathbf{X} \in \mathbf{R}^{m \times n}$
Class label Vector: $\mathbf{Y} \in \mathbf{N}^{n \times 1}$
Sparsity-controlling parameter: $\gamma \geq 0$
Number of principal components: k
Output: Loading matrix: $\mathbf{Z} \in \mathbf{R}^{m \times k}$

The differential expression matrix $\mathbf{E}$ is obtained via (3).
The total scatter matrix $\mathbf{S}_t$ is obtained according to (6).
$\mathbf{U}$, Δ and $\mathbf{V}$ are obtained via decomposing $\mathbf{S_t}$ by SVD.

$\mathbf{F} \Leftarrow \big(\mathbf{U}\Delta^{\theta}\big)^T$.

For j = 1 : k
Loop

$$t \Leftarrow \sum\nolimits_{i=1}^n f_i \cdot \big\lfloor \big|f_i^T x\big| - \gamma \big\rfloor \cdot \operatorname{sgn}\big(f_i^T x\big).$$
$$x_{k+1} \Leftarrow t/\|t\|_2.$$

Until a stopping criterion is satisfied.

$$z_{ij} \Leftarrow \operatorname{sgn}\big(f_i^T x\big) \big\lfloor \big|f_i^T x\big| - \gamma \big\rfloor / \sqrt{\sum\nolimits_{k=1}^n \big\lfloor \big|f_i^T x\big| - \gamma \big\rfloor^2} i = 1, \cdots, m.$$
$$q_j \Leftarrow \mathbf{F} z_j.$$
$$\mathbf{F}_{j+1} \Leftarrow \mathbf{F}_j - q_j z_j^T.$$

End

2.7 Identification of Differentially Expressed Genes

In CISCA, $\mathbf{Q}$ is the matrix of factor scores, and $\mathbf{Z}$ is the loading matrix of the principal components (PCs), which transforms the data matrix into factor scores. The data matrix $\mathbf{F}$, factor scores matrix $\mathbf{Q}$ and PCs $\mathbf{Z}$ are shown in Fig. 1.

Fig. 1 shows that the PCs $\mathbf{Z}$ give the coefficients of the linear combinations to compute the factors scores $\mathbf{Q}$. So the bigger the absolute value of the elements in PCs $\mathbf{Z}$, the more contribution for the factor scores matrix, the more important the corresponding gene in $\mathbf{F}$. So we can select the characteristic genes according to the PCs $\mathbf{Z}$.

Let

$$Z_i = [z_{1i}, z_{2i}, \cdots, z_{mi}]^T, i = 1, \cdots, k \tag{21}$$

denote the *i*th PC, the PCs can be given as the follows:

$$\mathbf{Z} = [Z_1, Z_2, \cdots, Z_k]. \tag{22}$$

As the absolute value of the *i*th row of PCs $\mathbf{Z}$ somewhat denotes the importance of the *i*th gene, we take the sum of all the entries'

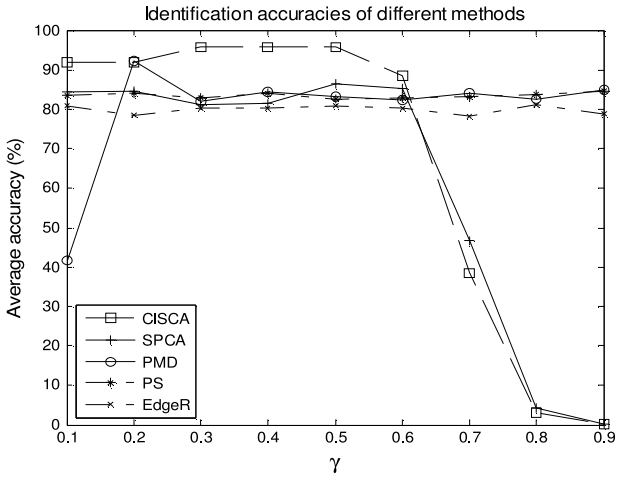

Fig. 2. An illustration of identification accuracies of these methods with different γ ($n = 6$). Note: PS and EdgeR are not sparse methods, which have nothing to do with γ.

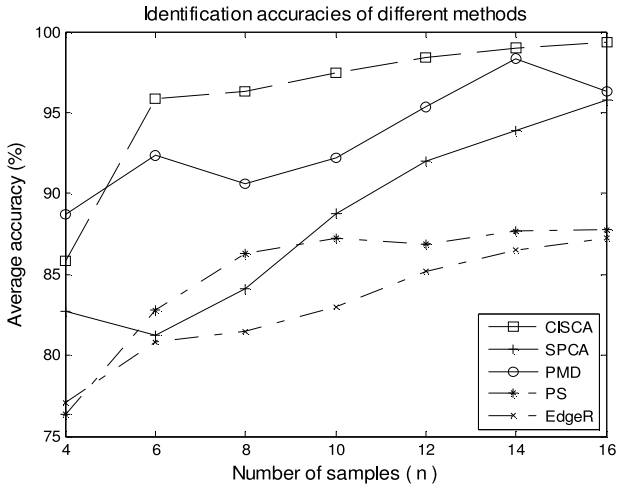

Fig. 4. Identification accuracies of the five methods on the simulation data set with different numbers of samples.

absolute value of the ith row as the evaluating vector $\mathbf{EV}$, which can be expressed as follows:

$$\mathbf{EV} = \left[\sum_{\mathbf{i=1}}^{\mathbf{k}} |\mathbf{z_{1i}}|, \sum_{\mathbf{i=1}}^{\mathbf{k}} |\mathbf{z_{2i}}|, \cdots, \sum_{\mathbf{i=1}}^{\mathbf{k}} |\mathbf{z_{mi}}|\right]^{\mathbf{T}}. \quad (23)$$

In particular, if the dimensionality of the gene data is m, then the $\mathbf{EV}$ has m entries. So we sort the evaluating vector $\mathbf{EV}$ in descending order and select the genes that have the first num largest entries as characteristic genes.

3 Results and Discussion

In this section, we evaluate the proposed method by applying it to identify the DE genes associated with a special biological process or biological function. Section 3.1 gives the results on simulation data. Results on real RNA-Seq data sets are given in Section 3.2. For comparison, we also use the PoissonSeq (PS) [7], SPCA [18], Penalized Matrix Decomposition [21] and EdgeR [29] methods to extract the features on these data sets.

3.1 Simulation Data

The simulation data are introduced in this Section 3.1.1. Then, the results are shown in Section 3.1.2.

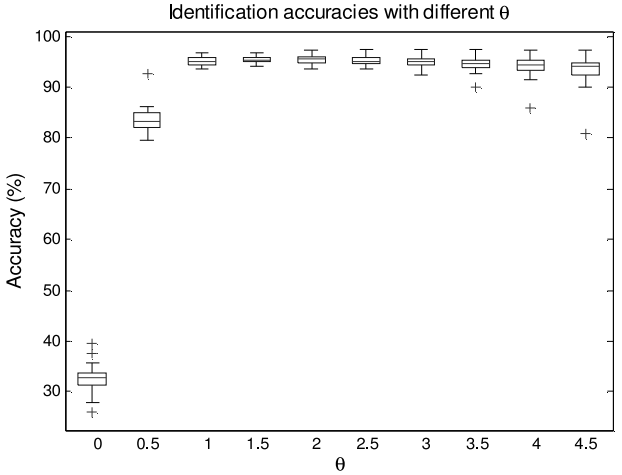

Fig. 3. An illustration of the identification accuracies of our method with different indices θ, when $n = 6$ and γ=0.3.

3.1.1 Data Source

We use (2) to generate the simulation data with $m = 20,000$ genes (roughly equal to the number of genes in the human genome) and $n = 4, 6, \cdots, 16$ samples. In the two-class case, we assign half of the samples to each class. To generate the mean counts of the samples $d_1, d_2, \cdots, d_n$, we let $\log d_j \tilde{\text{uniform}}(4, 6), \quad j = 1, \cdots, n$, which gives us between 1 and 8 million total counts per sample. To obtain the gene expression profile analogous to a real RNA-Seq dataset, we let $\beta_i = N_i/(\sum_{k=1}^{m} N_k/m)$, where $N_i, \quad i = 1, \cdots, m$ are the counts of all the genes in the Wang dataset [30]. We let $e_{ij} = y_j * \varphi_i$ in (2), where y_j is the label of each class. For 90 percent genes of non-differential expression, $\varphi_i = 0$, and 7 percent genes are up-regulated with $\varphi_i = 1$, and 3 percent genes are down-regulated with $\varphi_i = -1$. We randomly assign non-negative integer numbers to the indices of DE genes.

3.1.2 Results on Simulation Data

In this experiment, the DE genes are identified by using PS, PMD, SPCA, EdgeR and CISCA. Except for PS and EdgeR, other methods are sparse versions, whose sparsity-controlling parameter γ has influence on identification accuracy. According to the algorithm in [21], in the case of PMD, γ should be restricted to the ranges $[1/\sqrt{m}, 1)$. Here, we test γ in the interval $[0.1, 0.9]$ with $step = 0.1$. We iterate 30 times to randomly generate the simulation data. Fig. 2 shows the average identification accuracies of these methods with different γ while $n = 6$. From this figure, we note that when $\gamma \in [0.3, 0.5]$, our method (CISCA) can get the highest identification accuracy. When $\gamma \in [0.1, 0.6]$, SPCA can give the higher identification accuracy. However, when $\gamma > 0.7$, the identification accuracies by using both SPCA and CISCA are decreased, which may be caused by the too strong constraint. When $\gamma = 0.2$, PMD gives the highest identification accuracy. When $\gamma \geq 0.3$, the identification accuracies of PMD method are close to PS and EdgeR. So in the following experiments, γ is set to 0.3 for SPCA and CISCA. For PMD, γ is set to 0.2.

TABLE 1
An Overview of the Data Sets

Data sets	Number of samples	Number of classes	Number of reads
Maqc	14	2	71,970,164
Wang	22	17	223,929,919

TABLE 2
The GO Terms of Genes Identified by These Methods on Maqc Data Set

Rank	Name	TIA	CISCA		SPCA		PMD		PS		EdgeR	
			P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit
1	Genes with H3K27me3 in MEF cells.	589	1.32E-55	95	1.41E−54	94	1.84E−52	92	1.89E−21	57	7.99E−47	86
2	synaptic transmission	788	2.31E−38	96	4.59E−36	93	1.33E−30	86	4.17E−28	84	9.32E−18	66
3	Set 'H3K27 bound': genes mark in human embryonic stem cells, as identified by ChIP on chip	1,117	9.76E−34	98	5.12E−33	97	4.46E−28	90	1.64E−18	75	1.58E−32	96
4	Genes with H3K27me3 in MCV6 cells	434	5.78E−33	62	7.39E−35	64	4.65E−31	60	2.98E−13	38	6.71E−21	48
5	Human Brain Chen−Plotkin08 747genes	594	8.80E−32	70	8.80E−32	70	6.96E−31	69	3.79E−42	82	5.02E−16	49
6	cell−cell signaling	1,281	2.23E−31	112	2.71E−28	107	4.14E−25	102	2.97E−22	99	1.52E−19	91
7	Synapse	646	3.85E−25	70	4.76E−23	67	3.04E−18	60	4.77E−26	72	1.89E−09	44
8	transporter activity	1,220	1.27E−19	89	4.38E−19	88	2.26E−16	83	9.08E−13	77	4.51E−10	68
9	neuron part	1,130	3.71E−18	82	7.19E−21	87	1.89E−15	77	1.16E−32	107	2.73E-08	60

In this table, 'TIA(Term in Annotation)' denotes the number of genes associated with the term in global genome; 'Hit' denotes the number of genes associated with the term in query.

Then, the index θ in (12) can make a difference in the performance of our method, so we test the performances of our method with different indices θ in an interval $[0, 4.5]$ with $step = 0.5$. We iterate 30 times to randomly generate the simulation data with $n = 6$.

Fig. 3 shows the identification accuracies of our method with different indices θ, when $\gamma = 0.3$. From this figure, we can see that when $\theta = 0.0$, our method has very lower identification accuracy; when $\theta = 0.5$, the identification accuracy can reach above 80 percent; the identification accuracy has a peak at $\theta = 1.5$. It means that if the new matrix $\mathbf{F}$ is constructed by using $\left(\mathbf{U}\Delta^{1.5}\right)^T$, we can obtain the highest performance of our method. So θ is set to 1.5 in the following experiments.

The identification accuracy is closely related to the number of samples. Fig. 4 shows the average identification accuracies of these methods with different sample numbers. In this experiment, we set $\gamma = 0.3$ for SPCA and CISCA methods and set $\gamma = 0.2$ for PMD method. From this figure, it can be seen that the identification accuracies of all these methods are improved by increasing the sample numbers. While $n \geq 8$, PS method reaches a plateau in term of identification accuracy. Moreover, while $n \geq 6$, CISCA outperforms the other methods on identification accuracy and its identification accuracy can reach above 95 percent.

3.2 RNA-Seq Data Sets

Two publically available RNA-Seq data sets are used to evaluate our method, i.e., Maqc [31] and Wang [30]. An overview of the data sets can be found in Table 1. Typically, these data are assigned to some classes (usually, genes) based on their mapping to a common region of the target genome. By obtaining tens of millions of short reads from the transcript population of interest and mapping these reads to the genome, RNA-Seq produces digital (count) rather than analog signals [32]. Here, these RNA-Seq count data sets are downloaded from http://bowtie-bio.sf.net/recount [33]. For a comparison, 500 genes are selected by each of these methods.

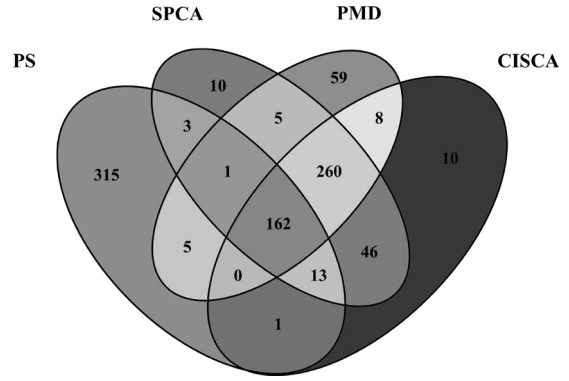

Fig. 5. Overlap among the sets of DE genes identified by the four methods on the Maqc data set.

3.2.1 Results on Maqc Data Set

The Maqc data set is composed of 14 samples which are derived from Ambion's human brain reference RNA and Stratagene's human universal reference RNA.

First of all, the Gene Ontology (GO) enrichment of functional annotation is checked by using ToppFun [34] which is publicly available at http://toppgene.cchmc.org/enrichment.jsp. We investigate the enrichment of functional annotations by inputting the 500 genes identified by these methods into the ToppFun, whose p-value is set to 0.01 and other parameters are used as default. Table 2 lists the closely related GO terms found by ToppFun.

In this table, there are 589 genes in the genome with the term of 'Genes with H3K27me3 in MEF cells'. SPCA, PMD, PS and EdgeR can identify 94, 92, 57 and 86 genes, respectively. At the same time, CISCA can identify 95 genes and has the lowest P-value (1.32E-55). Moreover, the SPCA can give the close P-value. This may be caused by the reason that our method (CISCA) is the generalization of SPCA. This table also lists some other terms with the most significance.

Next, we study the overlap among the sets of DE genes identified by using different methods. Fig. 5 shows the overlap among

TABLE 3
The GO Terms of Genes Shared by These Methods on Maqc Data Set

Rank	Name	P-value	Hit	TIA
1	Human Brain Chen-Plotkin08 747genes	6.03E-28	40	594
2	Genes with H3K27me3 in MEF cells.	5.80E-27	39	589
3	synaptic transmission	7.72E-24	44	788
4	neuron part	2.63E-18	44	1,130
5	cell-cell signaling	5.67E-18	47	1,281
6	Set 'H3K27 bound': genes mark in human embryonic stem cells, as identified by ChIP on chip.	6.51E-18	40	1,117
7	synapse part	2.75E-17	29	469
8	synapse	4.00E-17	33	646
9	neuron projection	4.88E-17	39	945

In this table, 'TIA(Term in Annotation)' denotes the number of genes associated with the term in global genome; 'Hit' denotes the number of genes associated with the term in query.

TABLE 4
The GO Terms of Genes Identified by These Methods on Wang Data Set

Rank	Name	TIA	CISCA		SPCA		PMD		EdgeR	
			P-value	Hit	P-value	Hit	P-value	Hit	P-value	Hit
1	Genes with H3K27me3 in MEF cells	589	3.51E-41	80	2.22E-33	72	5.31E-20	55	4.80E-10	39
2	Genes with H3K27me3 in MCV6 cells	434	1.40E-28	57	5.29E-19	46	2.75E-09	35	8.04E-06	25
3	Set 'Suz12 targets': genes identified by ChIP on chip as targets of the Polycomb protein SUZ12	1,038	5.52E-28	86	6.33E-18	71	1.65E-07	54	5.30E-16	67
4	Genes with H3K4me2 and trimethylation at K27 (H3K27me3) in brain	1,069	9.23E-28	87	1.94E-26	86	1.21E-15	68	1.39E-17	71
5	Genes with H3K27me3 in neural progenitor cells (NPC)	341	1.87E-26	49	7.96E-16	37	4.65E-10	32	6.03E-09	27
6	Human EmbryonicStemCell Xu09 1801genes	1,430	1.57E-25	98	1.98E-36	116	1.98E-21	92	4.36E-10	68
7	cell-cell signaling	1,281	4.65E-18	89	7.49E-17	88	1.35E-13	81	1.43E-05	17
8	neuron projection	945	4.27E-17	72	4.07E-10	59	2.79E-06	49	3.69E-03	38

In this table, 'TIA(Term in Annotation)' denotes the number of genes associated with the term in global genome; 'Hit' denotes the number of genes associated with the term in query.

the sets of DE genes. From this figure, we note that the DE genes identified by CISCA are to a large extent identified also by PS, SPCA and PMD. Here, only 10 genes identified by CISCA are not shared with other methods. In contrast, PS identifies a fair amount of 'unique' DE genes that are not shared with other methods (PS has 315 'unique' genes). Moreover, there are 481 genes shared by SPCA and CISCA, which may reflect CISCA is the generalization of SPCA and has similar performance with SPCA. There are 162 genes shared by all the four methods.

Finally, The 162 genes shared by the four methods are input into the GO tool. ToppFun's p-value is set to 0.01 and its other parameters are used as default. Table 3 lists the closely related GO terms found by ToppFun. There are 40 genes in the term of 'Human Brain Chen-Plotkin08 747genes'. The term has the lowest p-value (6.03E-28), so it is considered as the most probable enrichment item. Some other terms with the most significance are also listed in this table.

3.2.2 Results on Wang Data Set

The Wang data set contains 22 samples which are derived from the following tissues: adipose, brain, breast, cerebellum (Cerebellum #4), colon, heart, liver, lymph node, skeletal muscle and testes. Here, we assign the samples to 17 classes according to http://bowtie-bio.sourceforge.net/recount/phenotypeTables/wang_phenodata.txt. Since PS method is used in two-class case, here only SPCA, PMD, EdgeR and CISCA are compared.

We investigate the enrichment of functional annotations by inputting the 500 genes identified by these methods into the ToppFun, whose p-value is set to 0.01 and other parameters are used as default. Table 4 lists the closely related GO terms found by ToppFun.

In this table, there are 589 genes in the genome with the term of 'Genes with H3K27me3 in MEF cells'. SPCA, PMD and EdgeR can identify 72, 55 and 39 genes, respectively. At the same time, CISCA can identify 80 genes and has the lowest P-value (3.51E-41). The term has the lowest p-value (2.59E-40), so it is considered as the most probable enrichment item. Moreover, although our method (CISCA) is the generalization of SPCA, CISCA gives the superior identification perfomance on Wang data set. We note that 'cell-cell signaling' is also included in this table, which are consistent with the idea that these GO functional categories are likely to contribute to fundamental differences in the different human tissues [30]. This table also lists some other terms with the most significance.

4 Conclusion

In this paper, the novel method, CISCA, is proposed to improve the performance of identifying DE genes from RNA-Seq data. It introduces the class information via the total scatter matrix. This method takes advantage of class labels of samples, so it can improve the identification ability. By integrating the normalizing method of RNA-Seq data and matrix decomposition method, CISCA is suitable for analyzing the RNA-Seq data. Last but not least, on simulation and real RNA-Seq data, the results demonstrate that our method is effective.

In future, we will focus on the biological interpretation of the identification genes.

Acknowledgments

This work was supported in part by the NSFC under grant Nos. 61370163, 61373027, 61203376, and 61272339; the China Postdoctoral Science Foundation funded project, No. 2014M560264; the Shandong Provincial Natural Science Foundation, under grant Nos. ZR2013FL016, BS2014DX004, and ZR2012FM023; Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20140417172417174, JCYJ20130329151843309, CXZZ20140904154910774, and JCYJ20140904154645958). Y. Xu is the corresponding author.

References

[1] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, and D. Parkhomchuk, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.

[2] A. C. Frazee, B. Langmead, and J. T. Leek, "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinform.*, vol. 12, p. 449, 2011.

[3] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[4] A. Oshlack, M. D. Robinson, and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biol.*, vol. 11, no. 12, p. 220, 2010.

[5] K. D. Hansen, R. A. Irizarry, and W. Zhijin, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[6] M. Hu, Y. Zhu, J. M. Taylor, J. S. Liu, and Z. S. Qin, "Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq," *Bioinformatics*, vol. 28, no. 1, pp. 63–68, 2012.

[7] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani, "Normalization, testing, and false discovery rate estimation for RNA-sequencing data," *Biostatistics*, vol. 13, no. 3, pp. 523–538, 2012.

[8] S. Lee, P. E. Chugh, H. Shen, R. Eberle, and D. P. Dittmer, "Poisson factor models with applications to non-normalized microRNA profiling," *Bioinformatics*, vol. 29, no. 9, pp. 1105–1111, 2013.

[9] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statist. Sci.*, vol. 18, no. 1, pp. 71–103, 2003.

[10] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 3, pp. 754–764, May 2012.

[11] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[12] A. D'Addabbo, M. Papale, S. Di Paolo, S. Magaldi, R. Colella, V. d' Onofrio, A. Di Palma, E. Ranieri, L. Gesualdo, and N. Ancona, "SVD Based feature selection and sample classification of proteomic data," in *Proc. 12th Int. Conf. Knowl.-Based Intell. Inform. Eng. Syst.*, 2008, pp. 556–563.

[13] J. F. Pinto da Costa, H. Alonso, and L. Roque, "A weighted principal component analysis and its application to gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 1, pp. 246–252, Jan. 2011.
[14] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE Access*, vol. 3, 2015.
[15] H. Ji, X. Li, Q.-f. Wang, and Y. Ning, "Differential principal component analysis of ChIP-seq," *Proc. Nat Acad. Sci.*, vol. 110, no. 17, pp. 6789–6794, 2013.
[16] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard, "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, vol. 464, no. 7289, pp. 768–772, 2010.
[17] V. K. Singh, R. Garg, and M. Jain, "A global view of transcriptome dynamics during flower development in chickpea by deep sequencing," *Plant Biotechnol. J.*, vol. 11, pp. 691–701, 2013.
[18] M. Journée, Y. Nesterov, P. Richtarik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, 2010.
[19] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, 2006.
[20] H. Shen, and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *J. Multivariate Anal.*, vol. 99, no. 6, pp. 1015–1034, 2008.
[21] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.
[22] D. Lee, W. Lee, Y. Lee, and Y. Pawitan, "Super-sparse principal component analyses for high-throughput genomic data," *BMC Bioinform.*, vol. 11, no. 1, p. 296, 2010.
[23] J. Zhuang, M. Widschwendter, and A. E. Teschendorff, "A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform," *BMC Bioinform.*, vol. 13, no. 1, p. 59, 2012.
[24] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, and S. Fitzgerald, "Ensembl 2011," *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. D800–D806, 2011.
[25] F. Song, D. Zhang, Q. Chen, and J. Wang, "Face recognition based on a novel linear discriminant criterion," *Pattern Anal. Appl.*, vol. 10, no. 3, pp. 165–174, 2007.
[26] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[27] A. d' Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, 2007.
[28] L. Mackey, "Deflation methods for sparse pca," in *Proc. Adv. Neural Inform. Process. Syst.*, 2009, vol. 21, pp. 1017–1024.
[29] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
[30] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.
[31] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinform.*, vol. 11, no. 1, p. 94, 2010.
[32] H. Jiang, and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009.
[33] A. Frazee, B. Langmead, and J. Leek, "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinform.*, vol. 12, no. 1, p. 449, 2011.
[34] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Res.*, vol. 37, no. suppl 2, pp. W305–W311, 2009.