

Differential Expression Analysis on RNA-Seq Count Data Based on Penalized Matrix Decomposition

Jin-Xing Liu, Ying-Lian Gao, Yong Xu*, *Member, IEEE*, Chun-Hou Zheng, *Member, IEEE*, and Jane You

Abstract—With the development of deep sequencing, vast amounts of RNA-Seq data have been generated. It is crucial how to extract and interpret the meaningful information contained in deep sequencing data. In this paper, based on penalized matrix decomposition (PMD), a novel method, named PMDSeq, was proposed to analyze RNA-seq count data. Firstly, to obtain the differential expression matrix, the matrix of RNA-seq count data was normalized. Secondly, the differential expression matrix was decomposed into three factor matrices. By imposing appropriate constraint on factor matrices, the PMDSeq method can highlight the differentially expressed genes. Thirdly, the proposed method can identify the differentially expressed genes based on the scaled eigensamples. Finally, we used gene ontology tools to check these differentially expressed genes. The experimental results on simulation and three real RNA-seq count data sets demonstrated the effectiveness of our method.

Index Terms—Deep sequencing, differential expression analysis, gene selection, matrix decomposition, RNA-seq data.

I. INTRODUCTION

CHANGES IN transcription are the most important mechanisms of differentiation and regulation. Until recently, the transcriptional activities of a cell are measured by PCR [1] in the case of few genes, or microarrays [2]–[4] which are used to investigate the whole transcriptome of an organism or tissue. Both methods require an existing knowledge about the organism’s transcripts, either in the form of ESTs or a complete reference genome sequence for primer or probe design [5]. Moreover, when gene expression levels are very low or high,

Manuscript received December 30, 2012; revised October 02, 2013; accepted December 04, 2013. This work was supported in part by the NSF under grant Nos. 61370163, 61233011, 61332011 and 61272339; Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20120613153352732 and JCYJ20130329151843309); the Shandong Provincial Natural Science Foundation, under grant No. ZR2013FL016; the Foundation of Qufu Normal University, No. XJ200947. Asterisk indicates corresponding author.

J. X. Liu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055 China and with the College of Information and Communication Technology, Qufu Normal University, Rizhao, 276826 China (e-mail: sdcavell@126.com).

Y. L. Gao is with the Library of Qufu Normal University, Qufu Normal University, Rizhao, 276826 China (e-mail: yinliangao@126.com).

*Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, and with Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, 518055 China (e-mail: yongxu@ymail.com).

C. H. Zheng is with the College of Electrical Engineering and Automation, Anhui University, Hefei, 230039 China (e-mail: zhengch99@126.com).

J. You is with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csyjia@comp.polyu.edu.hk).

Digital Object Identifier 10.1109/TNB.2013.2296978

microarrays often lack sensitivity, or result in saturated signal. RNA-seq can overcome microarray associated problems with cross hybridization of similar sequences and allow of single nucleotide resolution, as well as reducing under-representation or the omission of low abundance sequences [6]. The advent of new deep sequencing technologies (also called second generation or next-generation sequencing methods) now allows us to study the transcriptome in unprecedented detail by directly sequencing the pool of expressed transcripts with high accuracy across many orders of expression magnitude [7], [8].

After obtaining tens of millions of short reads from the transcript population of interest by deep sequencing, RNA-seq produces digital (count) rather than analog signals by mapping these reads to a common region of the target genome [9]. For RNA-seq data, this read count has been found to be (to good approximation) linearly related to the abundance of the target transcript [10]. Then, the following issue posed to biologists is how to find the interesting information from the RNA-seq count data. The nature of the RNA-seq can result in different samples with dramatically different total number of sequence reads, so counts from each experiment should be “normalized” by the sequencing depth of that experiment before any comparison is made between experiments [11]–[13].

Problems of differential expression analysis include the identification of gene expression differences among different tissues, among diseased and healthy tissues, or among different species. For a given gene, we can consider an observed difference in read counts significant, when it is greater than what would be expected just due to natural random variation [14].

To decide whether the differential expression of a gene, the simplest and most common analysis approach is to compare the number of reads overlapping the exons in a gene between different biological conditions. The simplest approach solely considers every gene, so it neglects the correlations among genes.

To overcome the drawback, the methods of feature extraction can be used to obtain the interesting information [15]–[17]. In this paper, the competitive method, PMDSeq, produces shrunken estimates of differential expression, based on the penalized matrix decomposition (PMD). PMD has been shown to be useful for microarray analysis via imposing penalization on factor matrices [18]. However, for RNA-seq count data, we are not aware whether the penalization has been carefully examined or not. An additional danger is posed by sample outliers, which are more likely to be encountered in large datasets, and for which the behavior of the existing approaches is unknown [14]. Similarly, the presence of zero counts (e.g., all zeros in one of the compared experimental conditions) would produce missing values or spurious tests.

To the best of our knowledge, the PMD based method has never been proposed for analyzing RNA-seq data reported in the literature. In this article, we describe PMDSeq, an approach for differential expression analysis of RNA-seq transcriptional count data. Firstly, to obtain the differential expression matrix Δ , we normalize the matrix of RNA-seq count data. Then, PMDSeq assumes that the differential expression matrix Δ can be decomposed into three factor matrices, i.e., eigensamples \mathbf{U} , eigenpatterns \mathbf{V} and singular value matrix Σ , which is expressed as follows:

$$\Delta = \mathbf{U}\Sigma\mathbf{V}^T, \quad \mathbf{U}\mathbf{U}^T = \mathbf{I}_m, \quad \mathbf{V}\mathbf{V}^T = \mathbf{I}_n. \quad (1)$$

Generally, the matrices \mathbf{U} and \mathbf{V} are dense. That is, the eigensamples and eigenpatterns have connection with all genes and samples. For purposes of differential expression analysis, the sparse matrices of \mathbf{U} and \mathbf{V} are expected, so the penalization will be imposed on \mathbf{U} and/or \mathbf{V} . After the sparse eigensamples \mathbf{U} are obtained, we give an identification method of differentially expressed genes based on them. Finally, the differentially expressed genes are verified by using Gene Ontology tools.

Our work has the following contributions: firstly, it uses, for the first time, the method of PMD to analyze the RNA-seq count data. Secondly, it gives a practical method to identify the differentially expressed genes on the basis of sparse eigensamples. Thirdly, a large number of experiments are provided and the results demonstrate that our method is effective.

The remainder of this paper is organized as follows: in Section II, we introduce the methodology of PMDSeq. Section III gives the experimental results and discussion. Finally, we provide some concluding remarks in Section IV.

II. METHODOLOGY

In this section, we will provide the method of normalizing RNA-seq count data, the definition of penalized matrix decomposition (PMD), biological model of PMD and identification method of differentially expressed genes.

A. Normalizing RNA-Seq Count Data

Let \mathbf{X} denote an $m \times n$ matrix, which consists of RNA-seq count data corresponding to m genes in n samples, in general, $m \gg n$. In the case of RNA-seq count data, for the sample j , x_{ij} is the expression level of the number of reads overlapping gene i included in the Ensembl annotation of the given organism's genome [19]. We assume that $x_{ij} \sim \text{Poisson}(\mu_{ij})$, where the form of μ_{ij} is given by a log-linear model as follows:

$$\log \mu_{ij} = \log d_j + \log \beta_i + \delta_{ij}. \quad (2)$$

Here, d_j is the sequencing depth for sample j , and without loss of generality, we assume that $\sum_{j=1}^n d_j = 1$. β_i captures the non-differential expression of gene i , which can be calculated by $\beta_i = \sum_{j=1}^n x_{ij}$. δ_{ij} captures the differential expression of gene i in sample j .

In order to obtain δ_{ij} , the sequencing depth d_j is estimated by using the method in [11]. Then, the differential expression δ_{ij} of gene i in sample j can be calculated as follows:

$$\delta_{ij} = \log \mu_{ij} - \log d_j - \log \beta_i. \quad (3)$$

After obtaining the element δ_{ij} of differential expression matrix Δ , we analyze the matrix Δ by PMD method to identify the differentially expressed genes.

B. Definition of PMD

This subsection briefly introduces the PMD method proposed by Witten *et al.* [20]. Without loss of generality, let the row means of Δ be zero, singular value decomposition (SVD) of matrix Δ [21] can be formulated as (1). The PMD generalizes SVD via imposing additional constraints on \mathbf{U} and/or \mathbf{V} . The PMD can be formulated as the following optimization problem:

$$\begin{aligned} & \underset{\sigma, \mathbf{u}, \mathbf{v}}{\text{minimize}} \quad \frac{1}{2} \|\Delta - \sigma \mathbf{u}\mathbf{v}^T\|_F^2 \\ & \text{s.t.} \quad \|\mathbf{u}\|_2^2 = 1, \|\mathbf{v}\|_2^2 = 1, \Psi_1(\mathbf{u}) \leq \lambda_1, \Psi_2(\mathbf{v}) \leq \lambda_2, \sigma \geq 0, \end{aligned} \quad (4)$$

where \mathbf{u} is a column vector of \mathbf{U} , \mathbf{v} is a column vector of \mathbf{V} , σ is a diagonal element of Σ , $\|\cdot\|_F$ is the Frobenius norm, Ψ_1 and Ψ_2 are convex penalty functions that can take a variety of forms [20].

Let the rank of Δ be r and Σ be a diagonal matrix with diagonal elements σ , the following equation can be proved [20]:

$$\frac{1}{2} \|\Delta - \mathbf{U}\Sigma\mathbf{V}^T\|_F^2 = \frac{1}{2} \|\Delta\|_F^2 - \sum_{k=1}^r \mathbf{u}_k^T \Delta \mathbf{v}_k \sigma_k + \frac{1}{2} \sum_{k=1}^r \sigma_k^2. \quad (5)$$

Hence, while $r = 1$, we can see that problem (4) can be equivalent to the maximization problem:

$$\begin{aligned} & \underset{\mathbf{u}, \mathbf{v}}{\text{maximize}} \quad \mathbf{u}^T \Delta \mathbf{v} \\ & \text{s.t.} \quad \|\mathbf{u}\|_2^2 = 1, \|\mathbf{v}\|_2^2 = 1, \Psi_1(\mathbf{u}) \leq \lambda_1, \Psi_2(\mathbf{v}) \leq \lambda_2, \end{aligned} \quad (6)$$

and the σ satisfying (4) is $\sigma = \mathbf{u}^T \Delta \mathbf{v}$.

The objective function $\mathbf{u}^T \Delta \mathbf{v}$ in (6) is bilinear in terms of \mathbf{u} and \mathbf{v} , that is, with \mathbf{u} fixed, it is linear in terms of \mathbf{v} , and vice versa. The optimization problem in (6) can be converted into the biconvex optimization problem as follows [20]:

$$\begin{aligned} & \underset{\mathbf{u}, \mathbf{v}}{\text{maximize}} \quad \mathbf{u}^T \Delta \mathbf{v} \\ & \text{s.t.} \quad \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1, \Psi_1(\mathbf{u}) \leq \lambda_1, \Psi_2(\mathbf{v}) \leq \lambda_2. \end{aligned} \quad (7)$$

It turns out that the solution to (7) satisfies (6) provided that λ is chosen appropriately [20]. Equation (7) is called the rank-1 PMD, and it can be solved by the iterative algorithm.

To obtain multiple components of PMD, we can use deflation method to maximize the criterion in (7) repeatedly, each time using the residual obtained by subtracting the product of previous factors $\sigma \mathbf{u}\mathbf{v}$ from Δ , i.e., $\Delta^{k+1} \leftarrow \Delta^k - \sigma_k \mathbf{u}_k \mathbf{v}_k^T$. Without the Ψ_1 - and Ψ_2 -penalty constraints, it can be shown that the K -factor PMD algorithm leads to the rank- K SVD of Δ . The detailed algorithm of PMD can be found in [20].

C. Biological Model Based on PMD

As mentioned in Section II-B, the PMD algorithm decomposes the matrix Δ of differential expression into two base matrices \mathbf{U} and \mathbf{V} . \mathbf{U} and \mathbf{V} are the left and right singular vectors,

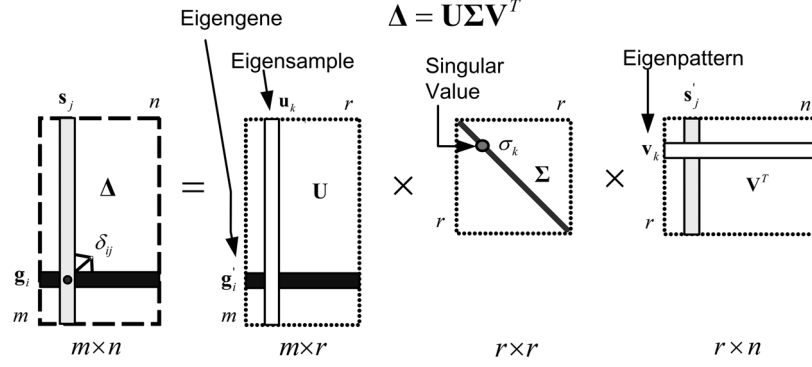


Fig. 1. Graphical description of PMD of a matrix Δ . \mathbf{U} and \mathbf{V} are the left and right singular vectors, respectively. Σ is a singular value matrix.

respectively. Following the convention [22], the left singular vectors $\{\mathbf{u}_k\}$, i.e., the columns of \mathbf{U} , are known as eigensamples and the rows of \mathbf{U} are known as eigengenes. The right singular vectors $\{\mathbf{v}_k\}$, i.e., the columns of \mathbf{V} are known as eigenpatterns. Eigensamples, eigengenes, eigenpatterns and other definitions are shown in Fig. 1.

The goal of differential expression analysis is to highlight genes that have significantly changed in abundance across experimental conditions. According to the definition of PMD, the interest signals in this case are the sample expression profile s_j . By (4), the PMD equation for s_j is

$$s_j = \sum_{k=1}^r v_{jk} \sigma_k \mathbf{u}_k, \quad j = 1, 2, \dots, n \quad (8)$$

which is a linear combination of the eigensamples $\{\mathbf{u}_k\}$.

By (4), the PMD equation for the gene transcriptional response g_i is

$$g_i = \sum_{k=1}^r u_{ik} \sigma_k \mathbf{v}_k, \quad i = 1, 2, \dots, m \quad (9)$$

which is a linear combination of the eigengenes $\{\mathbf{v}_k\}$.

Referring to the definition of SVD, we know that the left singular vectors span the space of the sample profiles $\{s_j\}$ and the right singular vectors span the differentially expressed space of the genes $\{g_i\}$ (see Fig. 1). So the left singular vectors $\{\mathbf{u}_k\}$ reflect the intensity of these interest signals. For differential expression analysis, $\{\mathbf{u}_k\}$ can be used as the identification basis of differentially expressed genes.

D. Identification of Differentially Expressed Genes

In order to reconstruct the original matrix, we need the eigensamples $\{\mathbf{u}_k\}$ which are m -dimensional vectors. By choosing appropriate penalty function Ψ_1 , the sparse vectors $\{\mathbf{u}_k\}$ can be obtained.

The nonzero entries in $\{\mathbf{u}_k\}$ can be positive or negative, which may reflect the up- or down-regulations of gene expression. Here, our goal is to identify differentially expressed genes, so we only consider the absolute values of nonzero entries in $\{\mathbf{u}_k\}$. Because eigensamples have different importance [17], the power of $\{\sigma_k\}$ for exponent value θ , $\{\sigma_k^\theta\}$ are used to weight the eigensamples, where $\theta \in [0, \infty)$. According to (8), the sample s_j can be represented by the scaled eigensamples $\{\sigma_k^\theta \mathbf{u}_k\}$. Then the absolute values of entries in the scaled

eigensamples $\{\sigma_k^\theta \mathbf{u}_k\}$ are utilized to identify differentially expressed genes.

As absolute values of the entries in row i of the scaled eigensamples $\{\sigma_k^\theta \mathbf{u}_k\}$ somewhat represent the importance of gene i , the absolute value sum of all the entries in row i is viewed as the evaluating element of gene i . In particular, if the dimensionality of the gene set is m , the evaluating vector (E) has m entries. The E can be formulated as follows:

$$E = [e_1 \quad e_2 \quad \dots \quad e_m]^T, \quad (10)$$

where $e_i = \sum_{j=1}^r |\sigma_j^\theta u_{ij}|$, $i = 1, \dots, m$.

Consequently, we sort the entries in E in descending order and obtain the new evaluating vector \tilde{E} . Without loss of generality, suppose that the first c entries in \tilde{E} are non-zero, that is,

$$\tilde{E} = \left[\tilde{e}_1, \dots, \tilde{e}_c, \underbrace{0, \dots, 0}_{m-c} \right]^T. \quad (11)$$

The idea of feature selection is as follows: if some elements of the evaluating vector are zero, the deletion of the associated input variables does not increase the reconstruction error of the original matrix for the remaining variables. Even the element is not zero, if its value is small enough, the deletion of the associated input variable does not affect very much for the reconstruction error. That is, the larger the entry in \tilde{E} is, the more influence on the reconstruction error is. So, the genes corresponding to the first $\text{num}(\text{num} \leq c)$ largest entries in \tilde{E} can be selected as differentially expressed genes.

E. Experimental Scheme of PMDSeq

In order to identify differentially expressed genes, our experimental scheme of PMDSeq is set as follows.

Firstly, to avoid performing log-transform on zeros, we add 1 to all the entries in count data matrix. To lessen the natural random variation, we filter out the genes that have too small counts in the count matrix by comparing the row mean value in count data matrix with a threshold.

Secondly, to obtain the differential expression matrix Δ , the count matrix is normalized according to (3). Then, the rows $\{g_i\}$ of matrix Δ are centered.

Thirdly, the real rank of differential expression matrix Δ is identified using the method in the Section III-B2.

Fourthly, the matrix Δ is decomposed into three factor matrices: \mathbf{U} , \mathbf{V} and Σ by using PMD method.

Fifthly, differentially expressed genes are identified on the basis of the evaluating vector \tilde{E} in (11).

Finally, these differentially expressed genes are analyzed by using gene ontology tools.

III. RESULTS AND DISCUSSION

This section shows the experimental results on simulation and RNA-seq count data. On RNA-seq data, the differentially expressed genes identified by PMDSeq will be verified using Gene Ontology (GO) tools.

A. Results on Simulation Data

1) *Simulation Data*: We simulate data with $m = 20000$ genes (roughly equal to the number of genes in the human genome) and $n = 16$ samples. To generate the mean counts of the samples d_1, d_2, \dots, d_n , we let $\log d_j \sim \text{uniform}(4, 6)$, $j = 1, \dots, n$, which gives us between 1 million and 8 million total counts per sample. To generate the gene expression profile which is analogous to a real RNA-seq dataset, we let $\beta_i = N_i / (\sum_{k=1}^m N_k / m)$, where N_i , $i = 1, \dots, m$ are the counts of all the genes in the Wang dataset [23]. In the two-class case, we assign half of the samples to each class. We let $\delta_{ij} = y_j * \gamma_i$, where y_j is the label of each class. For 90% genes of non-differential expression, $\gamma_i = 0$, and 7% genes are up-regulated with $\gamma_i = 1$, and 3% genes are down-regulated with $\gamma_i = -1$. The indices of differentially expressed genes are assigned by randomizing non-negative integer numbers. Finally, the simulation data are generated by using (2).

2) *Simulation Results*: In this paper, differentially expressed genes are identified by the scaled eigensamples $\{\sigma_k^\theta \mathbf{u}_k\}$, so we only impose constraint on \mathbf{u} , i.e., $\Psi_1(\mathbf{u}) \leq \lambda_1$, and don't impose constraint on \mathbf{v} . According to the algorithm in [20], generally speaking, the ranges of λ_1 should be restricted in $[1, \sqrt{m}]$. Let $\lambda = \lambda_1 / \sqrt{m}$, λ should be restricted to the ranges $[1/\sqrt{m}, 1]$. Here, we test λ in the ranges $[1/\sqrt{m}, 1]$ and θ in the ranges $[0, \infty)$. To verify the performance of PMDSeq, we iterate 30 times to randomly generate the simulation data. Table I lists the identification accuracies of our method on the simulation data with different λ and θ in terms of a percentage. As listed in Table I, while $\theta \geq 1.0$ and $\lambda \geq 0.2$, the identification accuracies can reach above 95%, so in the following experiments on RNA-seq data, we will set $\theta = 1.0$ and $\lambda = 0.3$.

B. Results on RNA-Seq Data

1) *Data Source*: Three publicly available RNA-seq datasets are used: Wang [23], Sultan [24] and Blekhman [25]. All sequence read data can be downloaded from the GEO database at NCBI under accession numbers GSE12946, GSE11892 and GSE17274 [26]. An overview of the RNA-seq data sets can be found in Table II. Here, we download the RNA-seq count data from <http://bowtie-bio.sf.net/recount> [27].

The Wang data set contains 22 samples; here we only use 9 samples which are derived from the following tissues: adipose, brain, breast, colon, heart, liver, lymph node, skeletal muscle and testes.

TABLE I
THE TEST ACCURACY WITH DIFFERENT θ AND λ

$\lambda \backslash \theta$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	83.99	90.88	51.01	35.70	38.33	46.61	51.48	51.78	52.93	52.97
	± 1.19	± 4.60	± 1.80	± 2.45	± 2.60	± 4.30	± 4.56	± 4.14	± 4.76	± 4.54
0.5	84.04	96.01	90.39	88.62	90.31	91.84	93.69	93.40	92.98	93.39
	± 2.11	± 2.87	± 1.26	± 1.25	± 1.35	± 1.28	± 1.48	± 1.34	± 1.40	± 1.82
1.0	84.16	96.19	95.59	95.33	95.87	95.55	96.12	96.02	96.33	96.34
	± 1.19	± 2.44	± 1.85	± 1.79	± 1.68	± 1.86	± 2.11	± 1.88	± 1.85	± 1.36
1.5	83.89	97.51	96.95	95.86	96.42	97.32	97.22	96.46	96.74	96.64
	± 1.33	± 1.96	± 1.82	± 2.19	± 2.14	± 1.64	± 1.78	± 1.85	± 2.00	± 2.09
2.0	83.24	96.94	97.39	96.18	96.01	96.93	96.92	96.80	96.79	95.91
	± 3.12	± 2.43	± 1.72	± 2.03	± 2.36	± 1.54	± 1.66	± 1.90	± 2.12	± 2.28
2.5	83.12	96.77	97.11	95.83	96.00	95.65	96.59	97.18	95.90	96.33
	± 4.45	± 2.57	± 1.85	± 2.15	± 2.15	± 2.21	± 1.95	± 1.81	± 2.11	± 2.15
3.0	83.89	97.43	96.08	96.04	96.68	96.67	96.46	95.39	96.66	96.34
	± 1.51	± 2.45	± 2.12	± 2.08	± 1.59	± 1.66	± 1.58	± 2.12	± 1.75	± 2.32
3.5	82.29	97.91	96.72	95.38	96.33	95.73	96.11	97.15	95.57	96.21
	± 5.05	± 1.51	± 1.82	± 1.94	± 2.09	± 1.67	± 1.93	± 1.49	± 2.02	± 1.90
4.0	82.54	97.29	96.60	95.71	95.97	96.13	95.69	96.12	95.83	95.96
	± 3.79	± 2.25	± 1.90	± 1.84	± 2.25	± 2.02	± 1.94	± 1.80	± 2.11	± 2.20
4.5	83.25	97.58	96.51	95.89	96.04	95.94	95.71	96.06	96.46	96.01
	± 3.52	± 1.98	± 1.82	± 2.01	± 1.83	± 1.82	± 1.69	± 1.86	± 1.91	± 1.71
5.0	83.14	97.14	95.67	95.74	95.89	95.79	95.63	96.05	95.75	95.67
	± 2.99	± 2.35	± 1.92	± 1.92	± 1.75	± 2.05	± 2.26	± 2.30	± 1.86	± 2.36

TABLE II
AN OVERVIEW OF THE RNA-SEQ DATA SETS

Data	Number of samples (original)	Number of samples used	Number of reads
Wang	22	9	223,929,919
Sultan	4	4	6,573,643
Blekhman	6	6	41,356,738

The Sultan data set consists of 4 samples which are extracted from Ramos B cells and human embryonic kidney (HEK) 293T tissues.

The Blekhman data set contains 6 samples which are used to study transcript levels in humans, chimpanzees and rhesus macaques, using liver RNA samples from male and female derived from each species.

2) *Parameters Selection*: The number of components, i.e., the real rank of differential expression matrix Δ , is the most important parameter of PMDSeq. Following the rules for deciding how many principal components of PCA, our strategy is given as follows. Firstly, the singular values (SVs) are obtained by SVD of the matrix Δ . Secondly, the differences among the SVs, i.e., $\text{diffSV}_k = \text{SV}_k - \text{SV}_{k+1}$, $k = 1, \dots, n - 1$, are calculated. Then, the number of the components is decided against the following criterion: when the $\{\text{diffSV}_k\}$ shows the first inflection point, k is selected as the real rank of the matrix Δ . Table III lists the real ranks of the three differential expression matrices selected by this strategy.

In this paper, differentially expressed genes are identified according to scaled eigensamples $\{\sigma_k^\theta \mathbf{u}_k\}$, so we only impose constraint on \mathbf{u} , i.e., $\Psi_1(\mathbf{u}) \leq \lambda_1$, and don't impose constraint on \mathbf{v} . According to Section III-A2, we let $\theta = 1.0$ and $\lambda = 0.3$. For purposes of simplifying comparison, we roughly select 1000 genes as differentially expressed ones.

TABLE III

THE REAL RANKS OF THE THREE DIFFERENTIAL EXPRESSION MATRICES

Data	Wang	Sultan	Blekhman
Real rank	5	2	3

TABLE IV

THE GO TERMS ASSOCIATED WITH THE FUNDAMENTAL DIFFERENCES ON WANG DATA SET

ID	Name	P-value	Term in Query	Term in Genome
GO:0035637	multicellular organismal signaling	2.73E-23	122	752
GO:0007268	synaptic transmission	1.18E-22	111	652
GO:0019226	transmission of nerve impulse	1.74E-22	119	736
GO:0007267	cell-cell signaling	2.02E-19	146	1100
GO:0050877	neurological system process	2.43E-13	143	1238
GO:0048878	chemical homeostasis	2.55E-10	103	840
GO:0030182	neuron differentiation	1.26E-07	108	992
GO:0044057	regulation of system process	1.11E-05	59	455
GO:0007417	central nervous system development	3.57E-05	77	688
GO:0048468	cell development	1.12E-04	133	1466

3) *Experimental Results*: In this subsection, we will show the experimental results on the real RNA-seq data sets.

a) *Experiments on Wang Data Set*: On Wang data set, our method is firstly used to identify differentially expressed genes. Then the genes of differential expression are input into Gene Ontology (GO) tool: ToppGene Suite [28], whose P-value cutoff is set to 0.01, other parameters are set to default values. The results of gene list enrichment analysis are given in supplementary 1 (Suppl.xls). Wang *et al.* argued that differentially expressed genes were enriched for GO functional categories including “developmental processes,” “cell communication,” “signal transduction,” and “regulation of metabolism” that were likely to contribute to fundamental differences in the biology of different human tissues [23]. Here, the terms associated with the above categories are listed in Table IV. From Table IV, we can see that these categories are included in our GO results with very lower P-values. For example, the P-value of “multicellular organismal signaling” is 2.73×10^{-23} . As shown in Table IV, our method can identify the differentially expressed genes closely related to the fundamental differences.

Table V shows the GO terms of coexpression associated with human different tissues. As shown in Table V, with very lower P-value, our method can identify the differentially expressed genes of specific-tissue, associated with brain, liver, breast, heart and colon. These GO results are consistent with the experimental tissues of Wang data set. Furthermore, the enrichment term of “Mesenchymal Stem Cells_Yamamoto 08_1252 genes” has the lowest P-value 2.02×10^{-51} , which may reflect the specific-tissue genes’ affinity with Stem-Cell ones.

b) *Experiments on Sultan Data Set*: Sultan data set only includes 4 samples derived from a human embryonic kidney and a B cell line. On Sultan data set, our method is firstly used to identify differentially expressed genes. Then the differentially expressed genes are input into ToppGene Suite [28], whose P-value cutoff is set to 0.01, other parameters are set to default values. The results of gene enrichment analysis are given in

TABLE V

THE COEXPRESSION TERMS CLOSELY RELATED TO DIFFERENT TISSUES ON WANG DATA SET

Name	P-value	Term in Query	Term in Genome
MesenchymalStemCells_Yamamoto08_1252gene	2.02E-51	161	853
EmbryonicStemCell_Xu09_1801genes	5.63E-43	199	1430
'H3K27 bound': genes in embryonic stem cells.	7.06E-40	168	1116
Human Brain_Chen-Plotkin08_747genes	4.35E-30	106	594
'Eed targets': in human embryonic stem cells.	6.22E-22	132	1062
Liver selective genes	5.37E-19	54	238
Genes up-regulated in IIDC relative to DCIS.	7.14E-18	64	346
Leukemia_Figueroa09_1035genes	1.65E-16	99	774
Genes up-regulated in the normal-like subtype of breast cancer.	1.81E-16	73	464
Up-regulated genes in AILT compared to normal T lymphocytes.	2.61E-14	44	202
Genes down-regulated in prostate cancer samples.	7.73E-13	67	466
Genes down-regulated in PTC compared to normal tissue.	7.69E-12	42	214
Down-regulated genes distinguishing between EGC and normal tissue samples.	1.93E-11	55	356
Human Colon_Graudens06_863genes	2.00E-07	74	702
Up-regulated genes in the left ventricle myocardium of patients with heart failure.	4.36E-05	21	103

TABLE VI

THE GO TERMS CLOSELY RELATED TO HEK AND B CELL LINE ON SULTAN DATA SET

Name	P-value	Term in Query	Term in Genome
Human StemCell_Cai06_1370genes	9.24E-29	156	1190
Genes up-regulated in the normal-like subtype of breast cancer.	4.88E-18	76	464
Human Sarcoma_Missiaglia10_176genes	7.76E-18	37	114
Human Leukemia_Lenz08_385genes	2.17E-15	61	349
Genes down-regulated in AIDS-related primary effusion lymphoma samples.	9.31E-15	25	58
Genes up-regulated in prostate cancer samples.	4.10E-13	52	292
Genes down-regulated in luminal-like breast cancer cell lines.	1.18E-12	66	453
Human Prostate_Wallace08_489genes	2.79E-12	60	392
Genes specifically up-regulated in UCC tumors.	7.51E-12	59	389
Genes up-regulated in IDC relative to DCIS.	3.59E-11	54	346
sequence-specific DNA binding	4.68E-11	85	683
regulation of developmental process	2.27E-10	147	1449
Genes down-regulated in multiple myeloma cell lines treated with both decitabineTSA.	2.86E-10	43	244
regulation of cell differentiation	5.78E-10	114	1023
regulation of multicellular organismal development	1.01E-09	120	1110

supplementary 2 (Suppl.xls). Table VI lists the significant terms.

From Table VI, we can see that many GO terms can be identified with very lower P-value, such as “DNA binding,” “regulation of developmental process,” “regulation of cell differentiation,” and so on. In addition, another three terms have considerable P-values, which are “Human StemCell_Cai06_1370genes,” “Genes up-regulated in the normal-like subtype of breast cancer,” and “Human Sarcoma_Missiaglia10_176genes.”

Finally, the GO enrichment has many terms about disease, for example, breast cancer, sarcoma, AIDS, leukemia, prostate cancer, etc. It is well known lymph nodes are garrisons of B, T and other immune cells. They act as filters or traps for foreign particles and are important in the proper functioning of the

TABLE VII
THE GO TERMS CLOSELY RELATED TO CONSERVED SEXUALLY DIMORPHIC
GENE REGULATION ON BLEKHMAN DATA SET

Name	P-value	Term in Query	Term in Genome
Genes up-regulated in pulpal tissue extracted from carious teeth.	1.13E-13	43	207
Genes down-regulated in the luminal B subtype of breast cancer.	6.02E-13	73	554
Human Leukemia_Verhaak05_568genes	4.97E-12	61	427
defense response	2.41E-09	115	1122
Genes up-regulated in basal subtype of breast cancer samples.	2.44E-09	72	636
Human Lymphoma_Leval07_644genes	3.23E-09	61	492
Immune_Kong10_456genes_ImmPort_Cytokines	5.21E-09	56	433
inflammatory response	2.08E-08	64	486
response to oxygen-containing compound	3.47E-08	89	809
response to external stimulus	3.55E-08	122	1269
response to wounding	1.69E-07	109	1111
response to lipid	4.78E-07	69	585
response to other organism	7.53E-07	69	591
locomotion	2.82E-06	115	1253
leukocyte migration	3.34E-06	39	251
immune response	9.56E-06	102	1087
cell-cell signaling	1.47E-04	99	1100
response to lipopolysaccharide	2.65E-04	32	213

immune system. So, many differentially expressed genes associated with tumor or disease are included in the GO terms.

c) *Experiments on Blekhman Data Set:* On Blekhman data set, our method is firstly used to identify differentially expressed genes. Then the differentially expressed genes are input into ToppGene Suite [28], whose P-value cutoff is set to 0.01, other parameters are set to default values. The results of gene enrichment analysis are given in supplementary 3 (Supp3.xls). Consequently, we focus on enriched categories at the top of ranked lists and only report qualitative results that are consistent with data from [25], which are listed in Table VII.

In Table VII, the two GO terms of the smallest P-value are “Genes up-regulated in pulpal tissue extracted from carious teeth” and “Genes down-regulated in the luminal B subtype of breast cancer.” As shown in Table VII, many GO terms are relative to metabolism and catabolism of lipids, such as “response to lipid,” etc. In addition, as Blekhman *et al.* argued in [25], some enrichments of genes involved in “immune response,” “defense response,” and so on. From Table VII, we can draw a conclusion that our method can identify the differentially expressed genes with expression patterns that are consistent with conserved sexually dimorphic gene regulation.

IV. CONCLUSION

In this paper, we proposed a novel method, PMDSeq, to identify the differentially expressed genes on RNA-seq count data. In our method, we firstly filter out the genes that have too small counts in the count matrix. Then to obtain the differential expression matrix, the matrix of count data was normalized. The differential expression matrix was decomposed into three factor matrices by PMD method. With the sparse constraint on the decomposition factor, we can identify the genes associated with the special biological progress or condition. Finally, these genes were analyzed by using the gene ontology tool. The experimental results on simulation and real RNA-seq data sets demonstrated that our method can effectively highlight the differen-

tially expressed genes, which verified that PMDSeq was a powerful tool for analysis of differential expression.

In future, we will focus on the biological interpretation of these differentially expressed genes.

REFERENCES

- [1] N. V. Demidenko, M. D. Logacheva, and A. A. Penin, “Selection and validation of reference genes for quantitative real-time PCR in buckwheat (*Fagopyrum esculentum*) based on transcriptome sequence data,” *PLOS One*, vol. 6, no. 5, May 12, 2011.
- [2] L. Wang and P. C. H. Li, “Microfluidic DNA microarray analysis: A review,” *Analytica Chimica Acta*, vol. 687, no. 1, pp. 12–27, 2011.
- [3] P. Maji and C. Das, “Relevant and significant supervised gene clusters for microarray cancer classification,” *IEEE Trans. NanoBiosci.*, vol. 11, no. 2, pp. 161–168, 2012.
- [4] L. Muresan, J. Jacak, E. P. Klement, J. Hesse, and G. Schutz, “Microarray analysis at single-molecule resolution,” *IEEE Trans. NanoBiosci.*, vol. 9, no. 1, pp. 51–58, 2010.
- [5] E. R. Mardis, “The impact of next-generation sequencing technology on genetics,” *Trends Genetics*, vol. 24, no. 3, pp. 133–141, 2008.
- [6] P. J. Hurd and C. J. Nelson, “Advantages of next-generation sequencing versus the microarray in epigenetic research,” *Briefings Funct. Genomics Proteomics*, vol. 8, no. 3, pp. 174–183, 2009.
- [7] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, 2008.
- [8] P. Tonner, V. Srinivasasainendra, S. Zhang, and D. Zhi, “Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data,” *BMC Genomics*, vol. 13, no. 1, p. 412, 2012.
- [9] H. Jiang and W. H. Wong, “Statistical inferences for isoform expression in RNA-seq,” *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biol.*, vol. 11, no. 10, p. R106, 2010.
- [15] S. Liu, R. Y. Patel, P. R. Daga, H. Liu, G. Fu, R. J. Doerksen, Y. Chen, and D. E. Wilkins, “Combined rule extraction and feature elimination in supervised classification,” *IEEE Trans. NanoBiosci.*, vol. 11, no. 3, pp. 228–236, 2012.
- [16] J. X. Liu, Y. Xu, C. H. Zheng, Y. Wang, and J. Y. Yang, “Characteristic gene selection via weighting principal components by singular values,” *PLOS One*, vol. 7, no. 7, p. e38873, 2012.
- [17] M. Lee, H. Shen, J. Z. Huang, and J. Marron, “Biclustering via sparse singular value decomposition,” *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010.
- [18] J. X. Liu, C. H. Zheng, and Y. Xu, “Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition,” *Comput. Biol. Med.*, vol. 42, no. 5, pp. 582–589, 2012.
- [19] 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.
- [20] 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.
- [21] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.
- [22] F. Liang, “Use of SVD-based probit transformation in clustering gene expression profiles,” *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 6355–6366, Aug. 15, 2007.
- [23] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtkova, 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.

- [24] 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.
- [25] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad, "Sex-specific and lineage-specific alternative splicing in primates," *Genome Res.*, vol. 20, no. 2, pp. 180–189, 2010.
- [26] NCBI, National Center for Biotechnology Information, Oct. 10, 2012 [Online]. Available: <http://www.ncbi.nlm.nih.gov/>
- [27] A. Frazee, B. Langmead, and J. Leek, "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, vol. 12, no. 1, p. 449, 2011.
- [28] 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.



Jin-Xing Liu received the B.S. degree in electronic information and electrical engineering from Shandong University, China, in 1993; the M.S. degree in control theory and control engineering from QuFu Normal University, China, in 2003; and the Ph.D. degree in computer simulation and control from the South China University of Technology in 2008.

He currently works as a Postdoctoral Fellow in the Shenzhen Graduate School, Harbin Institute of Technology, China. His research interests include pattern recognition and bioinformatics.



Ying-Lian Gao received her B.S. and M.S. degrees at QuFu Normal University, China, in 1997 and 2000, respectively.

Now she is currently with the Library of Qufu Normal University, China. Her current interests include data mining and pattern recognition.



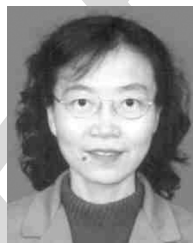
Yong Xu received his B.S. and M.S. degrees at Air Force Institute of Meteorology, China, in 1994 and 1997, respectively. He then received his Ph.D. degree in pattern recognition and intelligence system at the Nanjing University of Science and Technology, China, in 2005.

From May 2005 to April 2007, he worked at Shenzhen Graduate School, Harbin Institute of Technology (HIT), China, as a Postdoctoral Research Fellow. Now he is a professor at Shenzhen Graduate School, HIT. His current interests include pattern recognition, and machine learning.



Chun-Hou Zheng received the B.S. degree in physics education and the M.S. degree in control theory and control engineering from QuFu Normal University, China, in 1995 and 2001, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China in 2006.

He is a professor at the College of Electrical Engineering and Automation, Anhui University, Anhui, China. His research interests include pattern recognition and bioinformatics.



Jane You obtained her B.Eng. in electronic engineering from Xi'an Jiaotong University, China, in 1986 and the Ph.D. in computer science from La Trobe University, Australia, in 1992. She was a lecturer at the University of South Australia and senior lecturer at Griffith University from 1993 till 2002. She was awarded a French Foreign Ministry International Fellowship during 1993–1994.

Currently she is a full Professor at the Hong Kong Polytechnic University. Her research interests include image processing, pattern recognition, medical imaging, biometrics computing, multimedia systems, and data mining.