Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 916352, 13 pages
http://dx.doi.org/10.1155/2015/916352

Research Article

Dynamic Model for RNA-seq Data Analysis

Lerong Li and Momiao Xiong

Human Genetics Center, Division of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Correspondence should be addressed to Momiao Xiong; momiao.xiong@uth.tmc.edu

Received 4 December 2014; Accepted 16 February 2015

Academic Editor: Ernesto Picardi

By measuring messenger RNA levels for all genes in a sample, RNA-seq provides an attractive option to characterize the global changes in transcription. RNA-seq is becoming the widely used platform for gene expression profiling. However, real transcription signals in the RNA-seq data are confounded with measurement and sequencing errors and other random biological/technical variation. To extract biologically useful transcription process from the RNA-seq data, we propose to use the second ODE for modeling the RNA-seq data. We use differential principal analysis to develop statistical methods for estimation of location-varying coefficients of the ODE. We validate the accuracy of the ODE model to fit the RNA-seq data by prediction analysis and 5-fold cross validation. To further evaluate the performance of the ODE model for RNA-seq data analysis, we used the location-varying coefficients of the second ODE as features to classify the normal and tumor cells. We demonstrate that even using the ODE model for single gene we can achieve high classification accuracy. We also conduct response analysis to investigate how the transcription process responds to the perturbation of the external signals and identify dozens of genes that are related to cancer.

1. Introduction

Next-generation sequencing (NGS) technologies have revolutionized advances in the study of the transcriptome. The newly developed deep-sequencing technologies make it possible to acquire both quantitative and qualitative information regarding transcript biology. By measuring messenger RNA levels for all genes in a sample, RNA-seq provides an attractive option to characterize the global changes in transcription.

To generate RNA-seq data, the complete set of mRNAs are first extracted from an RNA sample and then shattered and reverse-transcribed into a library of cDNA fragments with adaptors attached. These short pieces of cDNA are amplified by polymerase chain reaction and sequenced by machine, producing millions of short reads. These reads are then mapped to a reference genome or reference transcript. The number of reads within a region of interest is used as a measure of abundance. The reads can also be assembled de novo without the genomic sequence to create a transcription map.

Compared to microarray which provides limited gene regulation information, RNA-seq offers a comprehensive picture of the transcriptome. RNA-seq has made a number of significant qualitative and quantitative improvements on gene expression analysis and provides multiple layers of resolutions and transcriptome complexity: the expression at exon, SNP, and positional level; splicing; posttranscriptional RNA editing across the entire gene; isoform and allele-specific expression [1].

Many advantages include strong concordance between platforms, higher sensitivity and dynamic range, lower technical variation and background signal, and high level of technical and biological reproducibility, and so on [2–5]. However, some limitations are inherent to next-generation sequencing technology. For example, the read coverage may not be homogeneous along the genome, and different samples may be sequenced at different levels of depth in the experiment. Also, although some genes may have a similar level of expression, longer genes are more likely to have more reads than short ones. Therefore, RNA-seq data must be normalized before any comparison of the counts can be made. Another consideration is that, in production of cDNA libraries, larger RNA must be fragmented into smaller pieces to be sequenced and different fragmentations may create bias towards different outcomes. Some other informatics challenges like the storage, transfer, and retrieval of large size data may bring additional errors [6, 7].

Expression variability measured by RNA-seq arises from three primary sources: (i) real biological differences in different experimental groups or conditions, (ii) measurement errors, and (iii) random biological and/or technical variation [1, 8]. The first type of variability is of real biological interest but is confounded with measurement and sequencing errors and other random biological/technical variation. How to appropriately take the latter two types of variability into account is essential issue in the RNA-seq data analysis.

The purpose of this paper is to borrow dynamic theory from engineering and use ordinary differential equation (ODE) for modelling the observed number of reads across the gene and unravelling the features of gene transcription [9]. To achieve this goal, we considered the number of reads or expression level at each position as a function of the genomic position and viewed the transcription process as a stochastic process of transcription along the genome. Instead of taking the derivative of expression level with respect to time, we calculated the expression level derivative with respect to genomic position. Specifically, we proposed a dynamic model for the variation of the transcription process along the genome. For each gene, we use a second order ODE with location-dependent coefficient to model that gene's transcription process. We develop statistical methods for estimation of the coefficient functions in the ODE based on principal differential analysis. Compared to the ODE model with constant coefficient to capture the stochastic variation feature of transcription process, the location-dependent coefficients are essential to account for the complicated stochastic process of gene regulation.

To examine the precision of the ODE for modeling the RNA-seq data, we split the samples into five groups and use 5-fold cross validation to evaluate the accuracy of predicting gene expression level across the gene using the ODE model.

To capture stochastic feature of gene regulation, we conduct the response analysis. The response analysis of transcriptional processes for each gene using its fitted differential equation can provide important aspects of transcription, including alternative splicing, alternative start and end of transcription, and alternative isoforms. To differentiate feature of gene regulation between normal and cancer tissue samples, we develop statistics to test for significant difference in the response of the gene regulation between the normal and cancer samples under the perturbation of external signals and perform genome-wide response analysis of gene regulation. Using the ODE model, we identified the genes that have a significantly different transcriptional process (both different magnitude and different patterns) and identified genes that showed significantly differential stochastic behaviors in response to environmental perturbations between normal and cancer samples.

To further explore application of the ODE for RNA-seq data analysis, we take the location-varying coefficients of the ODE as features and use FPCA as a tool for extraction of these features. The FPCA scores are used as features and the Lasso logistic regression is used as feature selection tool and classifier for distinguishing the cancer and normal samples.

The data suggest that the dynamic features of gene transcription captured by the coefficient functions can retrieve the original process information. Therefore, they naturally served as good candidate's features for clustering genes with similar transcription process. These groups of genes could share common biological function, chromosomal location, pathway, or regulation. The ODE for modeling the RNA-seq data has the potential to provide valuable information for understanding the mechanism of gene regulation and unraveling disease processes.

2. Materials and Methods

2.1. ODE Model with Varying Coefficients for RNA-seq Data. Assume that the expression of a gene is measured by the number of sequence reads mapped to this gene in the region $T = [a, b]$. Let t denote a genomic position, let $y(t)$ be observed gene expression level that was measured by the number of reads mapped to the genomic position, and let $x(t)$ be the hidden state that determined the gene expression level at the genomic position t. To model transcription process, the second order ordinary differential equation (ODE) with location-varying coefficients can be specified as follows:

$$L(x(t)) = \frac{d^2x(t)}{dt^2} + w_1(t)\frac{dx(t)}{dt} + w_0(t)x(t) = 0, \quad (1)$$

where $w_1(t)$ and $w_0(t)$ are weighting coefficients or parameters in the ODE. Its observations $y(t)$ often have measurement errors:

$$y(t) = x(t) + e(t), \quad (2)$$

where $e(t)$ is measurement error at the position t.

2.2. Estimation of Coefficient Functions in the ODE. Estimation of coefficient functions in the ODE consists of two steps. At the first step we estimate the states $\widehat{x}(t)$ from the observed number of reads $y(t)$ assuming that coefficient functions in the ODE are given. At the second step, we estimate the coefficient functions in the ODE, assuming that states $x(t)$ have been estimated.

Step 1. To estimate $x(t)$, we first expand the function $x(t)$ in terms of basis functions $\phi(t)$ and then estimate its expansion coefficients. Let $x_i(t)$ be the state variable at the genomic position t of the ith sample and let $y_i(t)$ be its observation ($i = 1, \ldots, n$). Then, $x_i(t)$ can be expanded as

$$x_i(t) = \sum_{j=1}^{K} c_{ij}\phi_j(t) = C_i^T\phi(t), \quad (3)$$

where $C_i = [c_{i1}, \ldots, c_{iK}]^T$ and $\phi(t) = [\phi_1(t), \ldots, \phi_K(t)]^T$.

Similarly, the parameters $w_1(t)$ and $w_0(t)$ can be expanded as

$$w_1(t) = \sum_{j=1}^{K} h_{1j}\phi_j(t) = h_1^T\phi(t),$$
$$w_0(t) = \sum_{j=1}^{K} h_{0j}\phi_j(t) = h_0^T\phi(t). \quad (4)$$

Substituting their expansions into (1), we obtain

$$L\left(x_i\left(t\right)\right) = C_i^T \Psi\left(t\right), \tag{5}$$

where $\Psi(t) = d^2\phi/dt^2 + G(t)h$, $G(t) = [(d\phi/dt)\phi^T(t), \phi(t)\phi^T(t)]$, and $h = [h_1^T, h_0^T]^T$. To smooth the estimated function $\widehat{x}(t)$, we impose the following penalty term:

$$\boldsymbol{\lambda} \int_T L\left(x_i\left(t\right)\right) L^T\left(x_i\left(t\right)\right) dt = \boldsymbol{\lambda} C_i^T J_{\phi h} C_i, \tag{6}$$

where $J_{\phi h} = \int_T \Psi(t)\Psi^T(t)dt$.

We estimate the state function $x(t)$ from the observation data $y(t)$ by minimizing the following objective function which consists of the sum of the squared errors between the observations and the estimated states and the penalty terms:

$$\begin{aligned}\sum_{i=1}^{n} \left\{ \sum_{j=1}^{\tau} \left[y_i\left(t_j\right) - x_i\left(t_j\right)\right]^2 + \boldsymbol{\lambda} \int_T L\left(x_i\left(t\right)\right) L^T\left(x_i\left(t\right)\right) dt \right\} \\ = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{\tau} \left[y_i\left(t_j\right) - x_i\left(t_j\right)\right] + \boldsymbol{\lambda} C_i^T J_{\phi h} C_i \right\}.\end{aligned} \tag{7}$$

Let

$$\begin{aligned} Y_i &= \left[y_i\left(t_1\right), \ldots, y_i\left(t_T\right)\right]^T, \qquad Y = \left[Y_1^T, \ldots, Y_n^T\right], \\ \widetilde{\phi} &= \left[\phi\left(t_1\right), \ldots, \phi\left(t_T\right)\right]^T, \qquad C = \left[C_1^T, \ldots, C_n^T\right]^T, \\ \Phi &= \operatorname{daig}\left(\widetilde{\phi}, \ldots, \widetilde{\phi}\right), \qquad J = \operatorname{diag}\left(J_{\varphi h}, \ldots, J_{\phi h}\right). \end{aligned} \tag{8}$$

Problem (7) can be rewritten in a matrix form:

$$\min_C \left(Y - \Phi C\right)^T \left(Y - \Phi C\right) + \boldsymbol{\lambda} C^T J C. \tag{9}$$

The least square estimators of the expansion coefficients are then given by

$$C = \left(\Phi^T \Phi + \boldsymbol{\lambda} J\right)^{-1} \Phi. \tag{10}$$

Step 2. Next we estimate the coefficient functions in the ODE. The coefficient functions in the ODE can be estimated by minimizing the following least squares objective function:

$$\min_h \mathrm{SSE}_p = \int_T L^T\left(\widehat{X}\left(t\right)\right) L\left(\widehat{X}\left(t\right)\right) dt, \tag{11}$$

where $L(\widehat{X}(t)) = [L(\widehat{x}_1(t)), \ldots, L(\widehat{x}_n(t))]^T$.

Since $L(x_i(t)) = C_i^T \Psi(t)$, the $L(x(t))$ can be expressed in terms of the estimated expansion coefficients as

$$L\left(\widehat{X}\left(t\right)\right) = C_* \Psi\left(t\right), \tag{12}$$

where the matrix C_* is defined as

$$C_* = \begin{bmatrix} C_1^T \\ \vdots \\ C_n^T \end{bmatrix}. \tag{13}$$

Therefore, problem (11) can be reduced as

$$\min_h \mathrm{SSE}_p = \int_T \Psi^T\left(t\right) C_*^T C_* \Psi\left(t\right) dt, \tag{14}$$

where the matrix C_* is estimated and hence fixed in minimization problem (14). Setting the partial derivative of SSE_p to be zero,

$$\frac{\partial \mathrm{SSE}_p}{\partial h} = \int_T G^T\left(t\right) C_*^T C_* \left[\frac{d^2\phi}{dt^2} + G\left(t\right) h\right] dt = 0. \tag{15}$$

Solving (15) for h, we obtain

$$h = -\left[\int_T G^T\left(t\right) C_*^T C_* G\left(t\right) dt\right]^{-1} \int_T G^T\left(t\right) C_*^T C_* \frac{d^2\phi}{dt^2} dt. \tag{16}$$

In summary, we iteratively determine the expansion coefficients of the state function $X(t)$ for fixed parameters in the ODE by (10) and estimate the coefficient functions in the ODE for fixed expansion coefficients by (16).

2.3. ODE for Classification. To illustrate that the ODE can be used as a useful tool for modeling the profile of the RNA-seq expression we will show that the ODE can capture all variation of gene expression across the gene and that the coefficient functions of the ODE are useful feature extraction of the RNA-seq data. The ODE can be used for classifying tumor and normal samples.

Since dimensions of the coefficient functions of the ODE are extremely high, the functional principal component analysis (FPCA) is used to reduce the dimensions of the coefficient functions of the ODE.

The FPCA tries to find the dominant direction of variation around an overall trend function [10, 11]. Each principal component is specified by the weight function $\beta(t)$, and the principal component scores of the individuals in the sample are defined as the inner product of weight function and functional curves $(w_0(t), w_1(t))$:

$$z = \int_T \beta\left(t\right) w\left(t\right) dt, \tag{17}$$

where, for convenience, we use $w(t)$ to denote either $w_1(t)$ or $w_0(t)$, that is, the coordinate value of functional curves at the direction of $\beta(t)$ with highest variability. By projecting the functional curves onto set of eigenfunctions, we can reduce the dimension to finite number, functional principal component scores.

Suppose that for the ith individual sample we obtain the functional principal component score:

$$\begin{aligned} z_{ij}^{(1)} &= \int_T \beta_j\left(t\right) w_{i1}\left(t\right) dt, \\ z_{ij}^{(0)} &= \int_T \beta_j\left(t\right) w_{i0}\left(t\right) dt, \end{aligned} \tag{18}$$

where $w_{i1}(t)$ and $w_{i0}(t)$ are the coefficient functions of the ODE for the ith individual sample and $\beta_j(t)$, $j = 1, \ldots, K$, are

a set of eigenfunctions (or principal component functions). The original functional curves can be reduced to a finite feature matrix:

$$Z = \begin{bmatrix} z_{11}^{(0)} & z_{11}^{(1)} & \cdots & z_{1K}^{(0)} & z_{1K}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ z_{n1}^{(0)} & z_{n1}^{(1)} & \cdots & z_{nK}^{(0)} & z_{nK}^{(1)} \end{bmatrix}, \tag{19}$$

where the K is the number of principal components selected to explain the total variability.

To improve classification accuracy we use the Lasso logistic regression as a classifier. In simple logistic regression, we use the logit link to relate the mean of response with the covariates of interest. Let $\mathbf{x}_i = [x_1, \ldots, x_p]^T$ be the vector of observed covariates for ith observation, and y_i is the corresponding response outcome. For simplicity, we consider binary cases where $y_i = 1$ or 0. The model is specified as the following posterior probability for ith observation [12]:

$$\pi_i(\mathbf{x}_i, \boldsymbol{\beta}) = \Pr(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)}, \tag{20}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is the covariate vector of interest and β_0 is the intercept term. And the joint log-likelihood of the N subjects is defined as

$$l(\beta) = \sum_{i=1}^{N} \log \pi_i(\mathbf{x}_i, \boldsymbol{\beta}) \tag{21}$$

which can be written as

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \{y_i \log \pi_i(\mathbf{x}_i, \boldsymbol{\beta}) + (1 - y_i) \log(1 - \pi_i(\mathbf{x}_i, \boldsymbol{\beta}))\} \\ &= \sum_{i=1}^{N} \left\{ y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right) - \log\left(1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}\right) \right\}. \end{aligned} \tag{22}$$

To estimate the parameter, we set its derivatives to zero and get the score equations

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} \mathbf{x}_i (y_i - \pi_i(\mathbf{x}_i, \beta)) = 0. \tag{23}$$

Since (23) is nonlinear equation in β, we usually use some iterative methods like Newton-Raphson algorithm to get the solution of β.

By adding an L_1 penalty to the joint log-likelihood in (23) we have the following constrained maximization equation:

$$l_c(\boldsymbol{\beta}) = \left\{ \sum_{i=1}^{N} \left\{ y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right) - \log\left(1 + e^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}\right) \right\} - \boldsymbol{\lambda} \sum_{j=1}^{p} |\beta_j| \right\}, \tag{24}$$

where $l_c(\boldsymbol{\beta})$ is the constrained log-likelihood and $\boldsymbol{\lambda}$ is tuning parameter to adjust the tradeoff between log-likelihood function and the size of penalty. Please note that, in Lasso, we usually do not penalize the intercept term and it is practically meaningful to standardize the covariates before optimization.

The L_1 penalty is not differentiable and also $\boldsymbol{\beta}$ is not linear solution of response $\mathbf{y}$. It is not trivial to get the score functions but we can still have a solution using nonlinear programming method [13]. The score functions for variables with nonzero coefficients have the form

$$\mathbf{x}_j^T (\mathbf{y} - \boldsymbol{\pi}) = \boldsymbol{\lambda} \cdot \operatorname{sign}(\beta_j). \tag{25}$$

Coordinate descent method is one efficient method to compute the Lasso solution. It fixes the penalty parameters $\boldsymbol{\lambda}$ and optimize over each parameter successively, while holding the others fixed at current values. R package *glmnet* [14] can efficiently fit the Lasso logistic regression with large N and p.

2.4. Numerical Solution to the ODE with Bounded Values. We use collocation Runge-Kutta method for the solution of boundary value problem of the ODE. The basic idea is to find a set of polynomials $p_n(x)$ of degree s which satisfies the problem over the interval $[x_{n-1}, x_n]$ for a set of points

$$x_{nj} = x_{n-1} + a_j h_n, \quad \text{where } j = 1, \ldots, s, \ n = 1, \ldots, N. \tag{26}$$

Note that, $0 < a_0 < a_1 < \cdots < a_s < 1$, they are distinct real numbers. Also the polynomial functions $p_n(x)$ are set to satisfy

$$\begin{aligned} &p_n(x_{n-1}) = y_{n-1}, \\ &p_n'(x_{nj}) = f(x_{nj}, p(x_{nj})), \\ &\qquad \text{where } j = 1, \ldots, s. \end{aligned} \tag{27}$$

The numerical approximation at x_n is given by

$$y_n = p_n(x_{n-1} + h_n). \tag{28}$$

R package *bvpSolve* implements the method for boundary value problem [15].

2.5. Response Analysis under Perturbation of External Signal. Gene regulatory properties are encoded in the parameter curves of the ODE modeling gene expressions. Testing significant difference in the parameter curves between two conditions can be used as a powerful tool to assess differential changing behaviors of the gene expression across the gene region between two conditions. Response analysis attempts to extract inherent features of the systems that capture and describe the behaviors of the system over genomic positions under different operating conditions and perturbation of external signals.

Let t denote a genomic region within the gene of interests and let $x(t)$ be the number of reads mapped to the genomic region. And the ODE model used to describe the expression profile is given as follows:

$$L(x(t)) = \frac{d^2 x(t)}{dt^2} + w_1(t) \frac{dx(t)}{dt} + w_0(t) x(t) = 0. \tag{29}$$

Suppose the $\widehat{w_1}(t)$ and $\widehat{w_0}(t)$ are estimated from the data. The response of a regulatory system depends on the input signals. Different signal will cause different responses. For simplicity, we consider unit-step signal forced on the system and then solve the responses of the original system between different groups using estimated parameters $\widehat{w_1}(t)$ and $\widehat{w_0}(t)$:

$$\frac{d^2x(t)}{dt^2} + \widehat{w_1}(t)\frac{dx(t)}{dt} + \widehat{w_0}(t)x(t) = U(t). \tag{30}$$

To solve the solution of the estimated ODE with unit-step force function $U(t)$, we have to use some numerical methods to approximate the solution $\hat{x}(t)$. We solved ODE numerically by considering two-point boundary value problems where boundary conditions are specified at both ends of the range of integration. We estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions.

Suppose $R(t) = [r_1(t), r_2(t), \ldots, r_{N_1}(t)]^T$ is a vector-valued function to represent response functional for all N_1 subjects in the normal group and $S(t) = [s_1(t), s_2(t), \ldots, s_{N_2}(t)]^T$ is response functional for N_2 subjects in cancer group. Therefore, we can construct a Hotelling T^2. Suppose that the response functions were expanded in terms of eigenfunctions $\phi_1(t), \ldots, \phi_K(t)$:

$$\begin{aligned} r_i(t) &= \sum_{j=1}^{K} \xi_{ij}\phi_j(t), \\ s_i(t) &= \sum_{j=1}^{K} \eta_{ij}\phi_j(t), \end{aligned} \tag{31}$$

where $\xi_{ij} = \int_T r_i(t)\phi_j(t)dt$ and $\eta_{ij} = \int_T s_i(t)\phi_j(t)dt$ and ξ_{ij} and η_{ij} are uncorrelated random variables with zero mean and variances $\boldsymbol{\lambda}_j$ with $\sum_j \boldsymbol{\lambda}_j < \infty$. Define the averages $\overline{\xi}_j$ and $\overline{\eta}_j$ of the principal component scores ξ_{ij} and η_{ij} in the normal and cancer group. Then we denote the average vector of scores in normal and cancer group by

$$\begin{aligned} \overline{\xi} &= \left[\overline{\xi_1}, \ldots \overline{\xi_k}\right]^T, \\ \overline{\eta} &= \left[\overline{\eta_1}, \ldots \overline{\eta_k}\right]^T, \end{aligned} \tag{32}$$

where $\overline{\xi_J} = (1/N_1)\sum_{i=1}^{N_1}\xi_{ij}$ and $\overline{\eta_J} = (1/N_2)\sum_{i=1}^{N_2}\eta_{ij}$, $j = 1, \ldots, k$.

The pooled covariance matrix is

$$S = \frac{1}{N_1 + N_2 - 2} \cdot \left(\sum_{i=1}^{N_1}\left(\xi_i - \overline{\xi}\right)\left(\xi_i - \overline{\xi}\right)^T + \sum_{i=1}^{N_2}\left(\eta_i - \overline{\eta}\right)\left(\eta_i - \overline{\eta}\right)^T\right), \tag{33}$$

where $\xi_i = [\xi_{i1}, \ldots, \xi_{ik}]^T$ and $\eta_i = [\eta_{i1}, \ldots, \eta_{ik}]^T$.

Let $\Lambda = (1/N_1 + 1/N_2)S$; then the Hotelling statistics can be written as

$$T^2 = \left(\overline{\xi} - \overline{\eta}\right)\Lambda^{-1} = \left(\overline{\xi} - \overline{\eta}\right)^T. \tag{34}$$

Under null of no difference in the response of the gene regulation between two groups, the statistics follows χ_k^2 distribution where k is the number of principle component scores.

3. Results

3.1. Dataset. We apply the proposed model to kidney renal clear cell carcinoma (KIRC) RNA-seq data, which is available from The Cancer Genome Atlas (TCGA) project (https://tcga-data.nci.nih.gov/tcga/). The RNA-seq data is available for 72 matched pairs of KIRC and normal samples. The maximum number of genomic positions where the expressions were measured by the number of reads passing quality control is 382,239,893 in the raw BAM file. And the total number of genes is 19,717.

Samtools and bedtools were applied to count number of reads for each base of the gene. Affected mapping reads were taken as the scale factor to normalize the reads for each individual. Hg19 human genome was taken as the reference.

Illumina paired-end RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq, and known genes from the UCSC genome browser, which was downloaded on August 19, 2010, August 8, 2010, and August 19, 2010, respectively. Reads mapped to junction regions were then repositioned back to the genome and were marked with "ZJ:Z" tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter were flagged with a custom script, and duplicated reads were flagged with Picard's MarkDuplicates.

In order to make the data comparable, we applied log transformation on the observed expression profiles. Some genomic position has zero counts and we intentionally add 1 to it and then it returns to be zero after log transformation. After that expression counts for most of genes are of the same scale. We also mapped the genes onto the interval [0, 100].

3.2. Evaluation of the ODE for Modeling RNA-seq Data. To evaluate the precision of the ODE for modeling the RNA-seq data, we first used the ODE to fit the RNA-seq data where the coefficient functions were estimated. Then, we used numerical collocation Runge-Kutta method to solve the fitted ODE. The solutions of the fitted ODE as a function of genomic position were then compared with the observed RNA-seq curves.

We estimated the varying-coefficient functions using the proposed model. The expression function for gene $X(t)$ was first estimated by spline smoothing with some initial penalty. We then update the penalty using the proposed second order ODE with varying-coefficient functions. We iterated between curve smoothing and ODE estimation until convergence was achieved. The smoothing parameters $\boldsymbol{\lambda}$ were chosen by cross validation process. By selecting the value of $\boldsymbol{\lambda}$, we trade off basis expansion fitting error and ODE solution filtering error.

Larger value of $\boldsymbol{\lambda}$ put more emphasis on the ODE penalty and the solution to ODE with estimated parameters is more likely to approximate the original data.

To validate the estimates of coefficients functions in the model, it is essential to compare the observed gene expression curve to the ODE solution with estimated coefficient functions. We solved ODE numerically by considering two-point boundary value problems where boundary conditions are specified at both ends of the range of integration. We estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions.

Figures 1(a) and 1(b) are fitted results of normal sample and cancer sample, respectively, for gene *CD74*. In these figures, the circles represent observed RNA-seq expression signal (green: normal; red: cancer); the blues lines are Fourier basis expansion to approximate the observed signal using weighted least square methods. The numbers of basis are chosen based on the length of genes and experimental adjustment to capture the important characteristics of gene expression curves. The dashed lines are estimated ODE solution using boundary value problem solver (R package *bvpSolve*). The ODE solutions approximate well the observed expression level of gene *CD74*, which show that estimated coefficient functions carry essential information of the original data. Once we have them, we can retrieve the original data very well.

To further evaluate the precision of the ODE for modeling RNA-seq data, we perform 5-fold cross validation prediction for gene *RPL29*. This method uses part of the available data to fit the model and estimate the parameters and uses the remaining data to test the model validity and estimate accuracy. We randomly split normal and cancer samples into five groups. From the estimation of parameters in the training samples, we solved the ODE with estimated coefficient functions to predict the expression curves of test samples. To be consistent, we estimated two initial values at both ends by evaluating the estimated smoothing expression curves at start and end positions in test samples. We also calculated the root mean square prediction error (RMSPE) for each folder to evaluate the performance of the prediction which is defined by

$$\text{RMSPE}_j = \frac{1}{N_j}\sum_{j=1}^{N_j}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{y}_i - y_i\right)^2}, \tag{35}$$

where y_i and $\widehat{y}_i$ are observed and predicted expression level; N is the number of genomic positions where the RNA-seq is observed for the gene and N_j is the number of subjects in the folder j.

Table 1 lists RMSPE in each folder for normal and cancer groups. The normal group has slightly better performance in terms of prediction on the test samples. But both prediction errors are relatively small.

Figures 2(a) and 2(b) are prediction results for selected samples in test set for gene *RPL29* in normal and cancer group, respectively. The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximations to the observed expression data. The dashed green lines indicate the predicted gene expression profile in the test set by solving the estimated ODE in the training examples. We can observe that all of the prediction can capture the overall shape and fluctuation in the data. Secondly, they can also predict the magnitude of expression value with decent accuracy. These are predicted very close to the observed expression profiles.

Table 1: RMSPE in each folder for normal and cancer groups.

Folder list	Normal RMSPE	Cancer RMSPE
1	0.23	0.97
2	0.31	0.94
3	0.24	0.79
4	0.33	0.70
5	0.30	0.73

Table 2: The average sensitivity, specificity, and accuracy of top 12 genes to classify normal and KIRC group over 5-fold cross validation.

Genes	Sensitivity	Specificity	Accuracy
RBBP8	0.903	0.958	0.931
ZFYVE16	0.903	0.958	0.931
LOC100129034	0.889	0.944	0.917
SLC44A2	0.931	0.903	0.917
TTC21B	0.903	0.931	0.917
C18orf56	0.958	0.861	0.910
KCNJ16	0.889	0.931	0.910
PFKP	0.917	0.903	0.910
TMCC1	0.903	0.903	0.903
CDK18	0.917	0.875	0.896
SEC61G	0.903	0.889	0.896
ST6GAL1	0.861	0.931	0.896

3.3. Classification Analysis. These data suggest that the estimated coefficient functions capture important features of expression curves. From the solution to estimated ODE, we can see the exceptional retrieval of original data. From the prediction performance in the test set, we can also get well-predicted curves by just proving two initial boundary data points. It is natural to consider them as features to classify phenotype categories.

We obtain two coefficient functions from one expression function. We can use FPCA to help us to reduce the dimension of features and to ease the computational effort. We first applied FPCA technique on two coefficient functions $w_0(t)$ and $w_1(t)$ separately; then we combined two groups of the selected functional principal component scores as aggregated features before we provided them to classifier. In the end, we applied Lasso logistic regression to help us select features and make prediction on the groups.

Table 2 lists top 12 genes to differentiate normal and KIRC group using 5-fold cross validations. We can see that using a single gene it can reach as high as 90% classification accuracy. These data strongly indicate that the ODE model

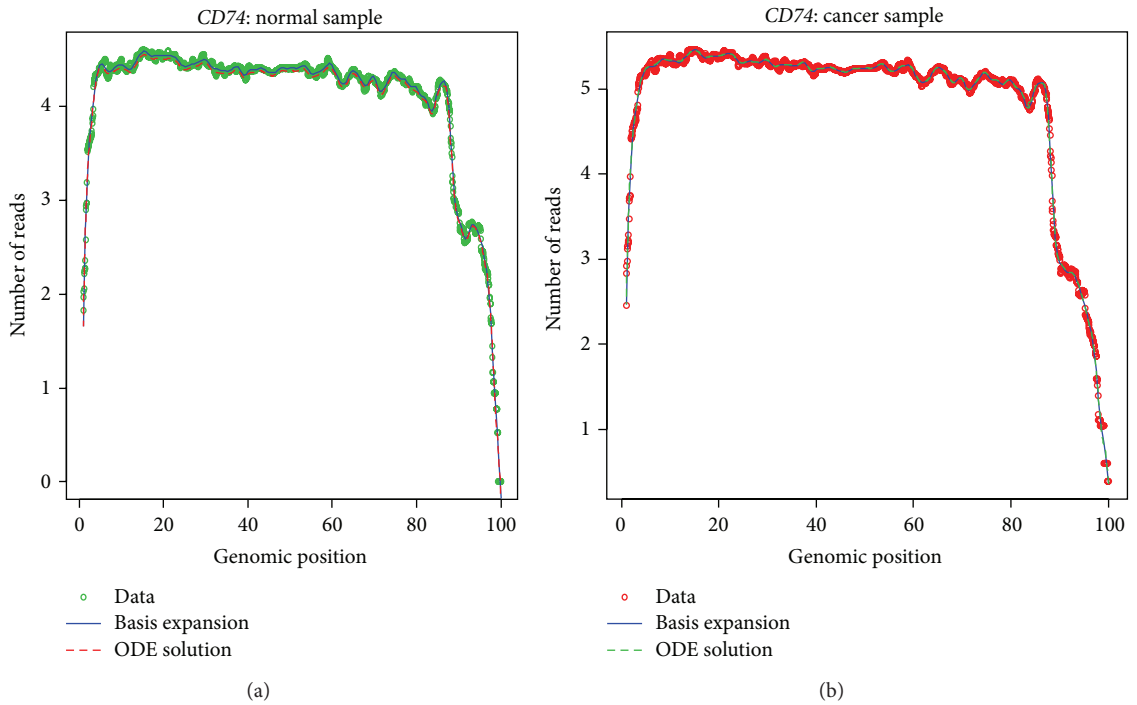


Figure 1: (a) Estimate of expression profiles for *CD74* by the ODE in a randomly selected normal sample. The green dotted points were observed expression levels, the blue solid lines are Fourier basis expansions, and the red dashed lines are numerical solution of ODE model. (b) Estimate of expression profiles for *CD74* by the ODE in a randomly selected tumor sample. The red dotted points were observed expression levels, the blue solid lines are Fourier basis expansions, and the green dashed lines are numerical solution of ODE model.

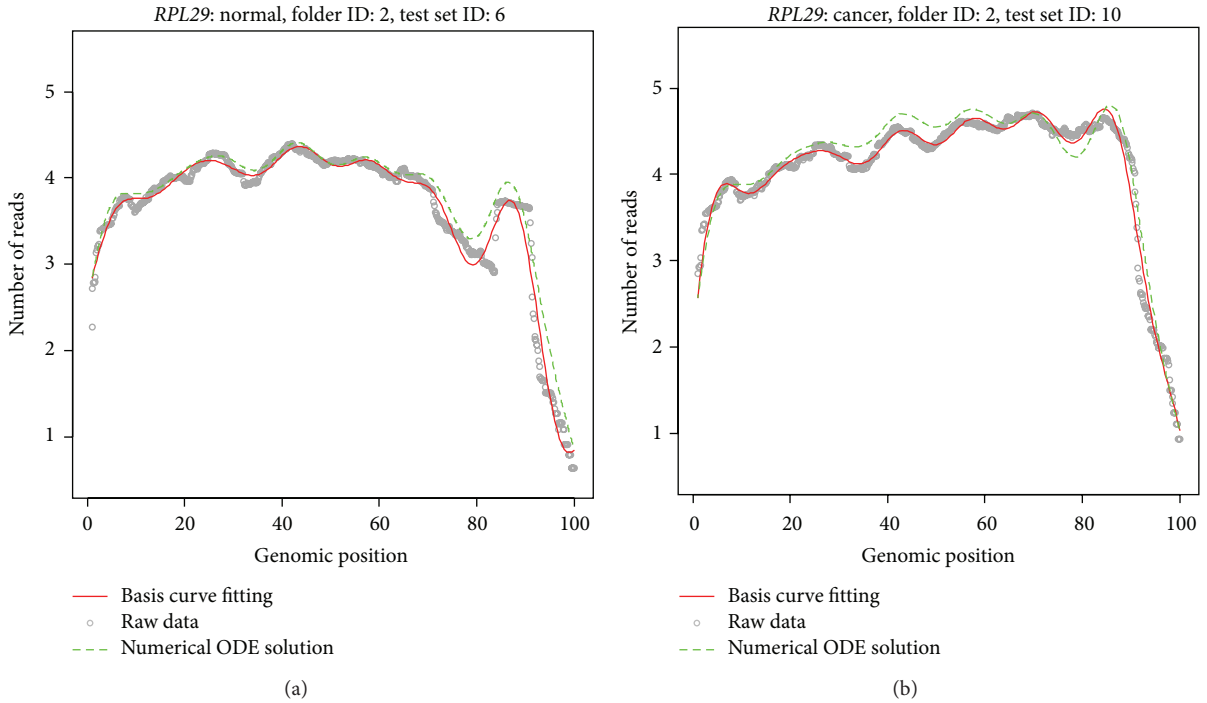


Figure 2: (a) Predicted expression curves for normal tissues for gene *RPL29*: The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximation to the observed expression data. The dashed green lines are predicted gene expression profile in the test set by solving the estimated ODE in the training examples. (b) Predicted expression curves for tumor tissues for gene *RPL29*: The gray dot is observed expression profile, and the solid red lines are Fourier basis expansion approximation to the observed expression data. The dashed green lines are predicted gene expression profile in the test set by solving the estimated ODE in the training examples.

effectively captured the inherent features of RNA-seq expression profile. We also evaluated the performance of classification result using sensitivity, specificity, and accuracy. Sensitivity is defined as the percentage of cancer tissues correctly classified as cancer. Specificity is defined as the percentage of the normal tissues correctly classified as normal. The classification accuracy is defined as the percentage of the correctly classified normal and cancer tissues. The classification results can reach as high as 99% if we use all these 12 genes together as predictors.

3.4. Genome-Scale Clustering Analysis. In this section, we continue to use the estimated coefficient functions as features to cluster genes expression data to study the genome-wide transcriptome. By grouping genes with similar patterns of expression profiles, cluster analysis can provide insight into gene functions and biological process. It also gives a simple way of determining the functions of many genes for which information is not available, as genes with the same functions may share expression profiles. We assume the coefficient functions in ODE model help to define these patterns in the dynamic regulation process and give us clues to functional discovery and pattern grouping.

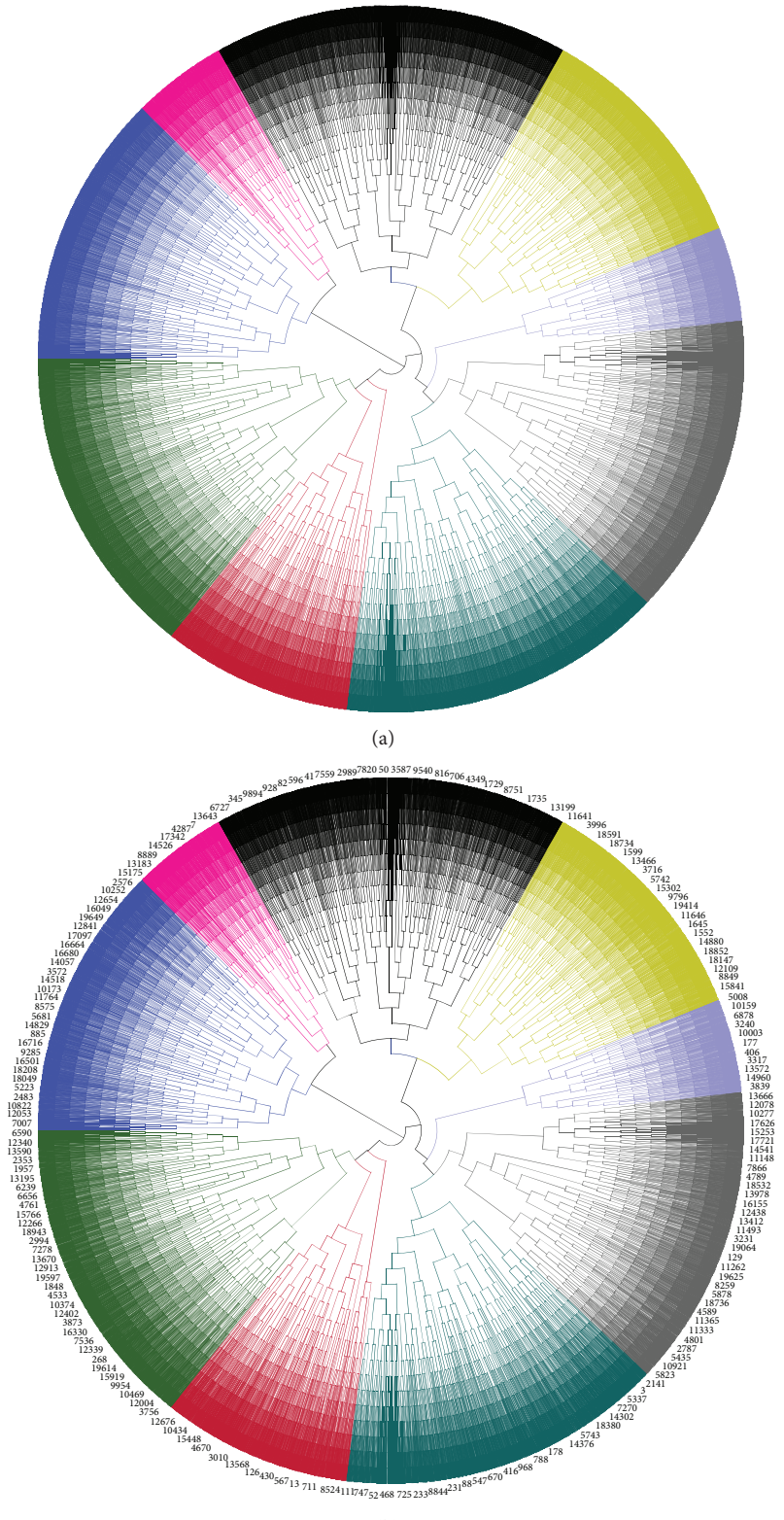
After we derive the feature matrix for all the genes from dimension reduction using FPCA, we merely need to adopt a metric definition which is used as a measure of similarity in the behavior of two genes. To calculate the distance matrix we used Euclidean distance and correlation matrix. This method computes a dendrogram that combines all genes in a single tree.

A total of 19717 genes were clustered into 9 groups according to the cluster analysis (Figures 3(a) and 3(b)). The functional principal component scores from coefficient functions in ODE model were used as significant features to define these patterns in the dynamic regulation process. The function annotation for each cluster was as follows.

The principle functions of the genes in the first group are mainly associated with oxidoreductase activity, ligase activity, dehydrogenase (NAD) activity, and related metabolic process. The detailed functions include aldehyde dehydrogenase (NAD) activity, translational initiation, mediator complex, MHC protein complex, mitochondrial membrane part, ion transmembrane transport, respiratory chain complex I, proton-transporting ATP synthase complex, proton-transporting two-sector ATPase complex, proton-transporting domain, NADH dehydrogenase (quinone) activity, oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, positive regulation of protein ubiquitination, response to unfolded protein, heterocycle metabolic process, protein modification process, mitochondrial ATP synthesis coupled proton transport, glycerolipid metabolic process, macromolecule modification, proton-transporting two-sector ATPase complex, catalytic domain, regulation of translational initiation, oxidoreductase activity, acting on the aldehyde or oxo group of donors, RNA polymerase II transcription mediator activity, heme binding, positive regulation of ligase activity, negative regulation of ligase activity, negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle, positive regulation of ubiquitin-protein ligase activity, negative regulation of ubiquitin-protein ligase activity, transferase activity, glycerolipid biosynthetic process, amine transport, phosphoinositide metabolic process, carboxylic acid transport, hormone binding, eukaryotic translation initiation factor 3 complex, glycerophospholipid metabolic process, helicase activity, response to protein stimulus, lipid biosynthetic process, phosphoinositide biosynthetic process, aldehyde dehydrogenase [NAD(P)+] activity, proton-transporting ATP synthase complex, coupling factor F(o), cytosolic part, nucleobase, nucleoside and nucleotide metabolic process, proton-transporting ATPase activity, rotational mechanism, phospholipid metabolic process, phosphorus metabolic process, phosphate metabolic process, hydrogen-exporting ATPase activity, phosphorylative mechanism, proton-transporting V-type ATPase complex, MHC class II protein complex, collagen, positive regulation of protein modification process, and posttranslational protein modification.

The principle functions of the genes in the second group are mainly associated with hydratase activity, cation transmembrane transporter activity and hydrolase activity, and related metabolic process. The detailed functions include NAD or NADH binding, peroxisomal membrane, microbody membrane, aconitate hydratase activity, 4 iron, 4 sulfur cluster binding, regulation of vesicle-mediated transport, lactate dehydrogenase activity, L-lactate dehydrogenase activity, long-chain fatty acid-CoA ligase activity, fatty acid ligase activity, homophilic cell adhesion, tight junction, cation transmembrane transporter activity, occluding junction, kinesin complex, microbody part, peroxisomal part, actin filament binding, hydrolase activity, hydrolyzing O-glycosyl compounds, hydrolase activity, and acting on glycosyl bonds.

The principle functions of the genes in the third group are mainly associated with monooxygenase activity, receptor activity, electron carrier activity, sodium ion transmembrane transporter activity, and related metabolic process The detailed functions include nucleoside binding, purine nucleoside binding, monooxygenase activity, receptor activity, protein binding, ATP-binding, DNA packaging, chromatin assembly or disassembly, electron carrier activity, sodium ion transmembrane transporter activity, actin cytoskeleton, RNA metabolic process, adenyl nucleotide binding, GTPase regulator activity, regulation of lipid transport, negative regulation of lipid transport, adenyl ribonucleotide binding, protein-DNA complex, very-low-density lipoprotein particle, triglyceride-rich lipoprotein particle, cellular nitrogen compound metabolic process, cellular macromolecule biosynthetic process, nucleosome organization, chylomicron, organelle, intracellular organelle, cellular biosynthetic process, keratin filament, regulation of transcription, regulation of biological process, regulation of cellular process, regulation of nitrogen compound metabolic process, regulation of RNA metabolic process, regulation of macromolecule metabolic process, nucleoside-triphosphatase regulator activity, biological regulation, DNA conformation change, and regulation of primary metabolic process.



(a)

(b)

Figure 3: (a) Circular phylogram tree of 19717 genes that were clustered into nine groups by Dendroscope 3.2.10. (b) Detailed circular phylogram tree of 19717 genes that were clustered into nine groups by Dendroscope 3.2.10.

The principle functions of the genes in the fourth group are mainly associated with acyl-CoA thioesterase activity, oxidoreductase activity, phosphatase activity, and related metabolic process. The detailed functions include organellar small ribosomal subunit, organellar large ribosomal subunit, phospholipid-translocating ATPase activity, glutathione transferase activity, receptor signaling protein serine/threonine kinase activity, transmembrane receptor activity, inward rectifier potassium channel activity, organic acid transmembrane transporter activity, mitochondrial matrix, mitochondrial large ribosomal subunit, mitochondrial small ribosomal subunit, cytosol, translation, translational elongation, cell surface receptor linked signaling pathway, large ribosomal subunit, small ribosomal subunit, integral to membrane, acyl-CoA thioesterase activity, oxidoreductase activity, acting on NADH or NADPH, phosphatase activity, cytosolic ribosome, signaling process, signal transmission, intrinsic to membrane, negative regulation of protein ubiquitination, cullin-RING ubiquitin ligase complex, and mitochondrial lumen.

The principle functions of the genes in the fifth group are mainly associated with cell projection part, microtubule associated complex, motor activity, microtubule, axoneme, microtubule-based process, microtubule-based movement, microtubule cytoskeleton, dynein complex, cytoskeletal part, cilium, macromolecular complex, cilium axoneme, cell projection, protein complex, cilium part, pyrophosphatase activity, hydrolase activity, acting on acid anhydrides, hydrolase activity, acting on acid anhydrides, phosphorus-containing anhydrides, and nucleoside-triphosphatase activity.

The principle functions of the genes in the sixth group are mainly associated with intracellular signal transduction, cholesterol efflux, UDP-galactosyltransferase activity, and histone demethylase activity.

The principle functions of the genes in the seventh group are mainly associated with ATP-binding cassette (ABC) transporter complex, JNK cascade, ATP-dependent peptidase activity.

The principle functions of the genes in the eighth group are mainly associated with glutamate receptor activity, ATPase activity, cytoskeletal protein binding, and myosin filament.

The principle functions of the genes in the ninth group are mainly associated with adrenoceptor activity, inhibition of adenylate cyclase activity by G-protein signaling pathway, adenosine deaminase activity, hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, cyclic amidines, deaminase activity, adenylate cyclase activity, activation of protein kinase A activity, alpha-adrenergic receptor activity, adrenergic receptor binding, epinephrine binding, regulation of norepinephrine secretion, norepinephrine transport, positive regulation of blood pressure, norepinephrine secretion, oxidoreductase activity, acting on CH-OH group of donors, oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor, and delayed rectifier potassium channel activity.

Since the ODE model use the dynamic information of gene expression profiles and we consider genes with similar expression profiles will share common biological functions. Based on the cluster analysis, the genes grouped together have similar pattern of expression share common biological function. It directs us a way to find the functions of many genes for which information is unknown by looking the genes in the same group.

Table 3: Genes with significant difference in response behavior between normal and tumor samples.

ABHD10	*MFSD1*	*SDR39U1*	*ATP6V1D*	*OXA1L*	*SEC31A*
BST2	*PACSIN2*	*SMCR8*	*CD74*	*PGAM1P5*	*SSR2*
DAP3	*PIK3CB*	*TM9SF2*	*DHX40*	*PITRM1*	*UBXN6*
EDF1	*POLR2B*	*UQCRC2*	*HLA-DMB*	*PSAP*	*VKORC1*
HSPA9	*PSMB10*	*ZNF710*	*ISYNA1*	*MAT2A*	*PSMB7*
PSMC4					

3.5. Response Analysis of Gene Regulation. The expression level of a gene measured by sequencing can be viewed as a curve or function of genomic position. The gene expression will vary across the gene region. If we treat time and space position as the same argument, all theories and methods of dynamic system can be applied to RNA-seq data analysis. The dynamic behavior of a system is encoded in the temporal evolution of its states or in the genomic location evolution of the gene expression in our problem. Therefore, borrowing dynamic theory, we can study the location-dependent variation of gene expression under the perturbation of the external signals. The transient response of the dynamic systems is an important property of the system itself. It can be used to quantify the space domain characteristics of the gene regulation system responding to the disturbance of environments. Our goal is to investigate how the gene expression level at each genomic position varies in response to the external perturbation and whether this will affect the function of cell.

We conducted response analysis of 19,717 genes under unit-step signal perturbation. We used the Hoteling T^2 statistic that was described in Section 2.5 to identify 31 genes that showed significant difference in the response property. The names of 31 genes with significant difference in response property were summarized in Table 3. In a few cases, the matrix Λ may be singular; we can use penalized method or generalized inverse to estimate Λ^{-1}. However, this will inflate the false positive rates.

We present Figures 4(a)–4(d) showing the average expression curves, unit-step response curves, and the coefficient curves of the ODE of gene *CD74*, respectively. We observed that gene *CD74* not only showed significant difference in gene expression and coefficient curves of the ODE but also demonstrated strong difference in the unit-step response. The changing point of gene expression curve and unit-step response curve occurred between 11b and 12a where a splicing site is located. It was reported that CD74 played critical role in cancer cell tumorigenesis [16] and downregulation of CD74 inhibits growth and invasion in clear cell renal cell carcinoma [17].

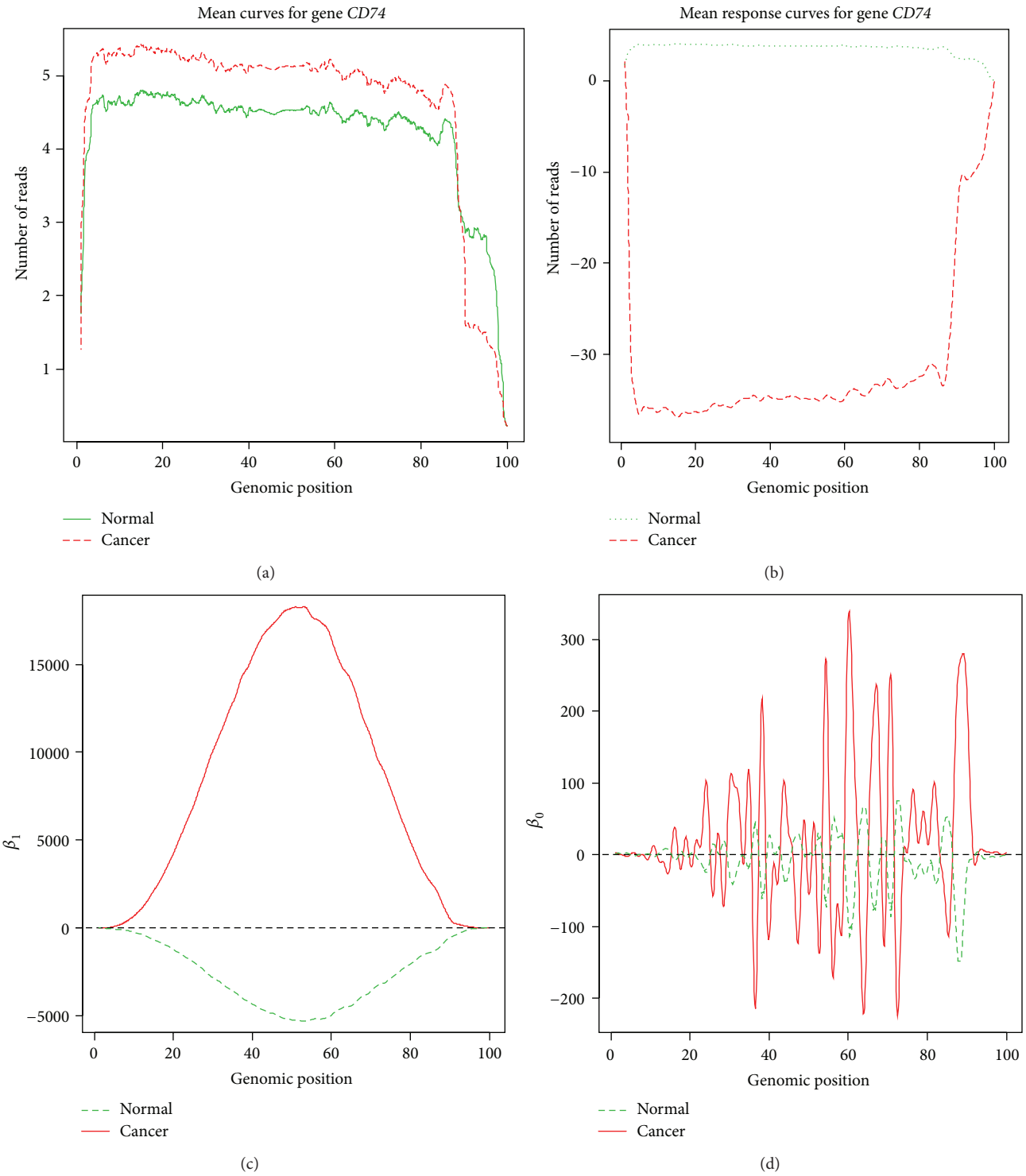


FIGURE 4: (a) Average expression curves of gene *CD74* in normal and tumor samples. (b) Average unit-step response curves of gene *CD74* in normal and tumor samples. (c) Average coefficient curves of the ODE for gene *CD74*. (d) Average coefficient curves of the ODE for gene *CD74*.

Transient response is one of dynamic properties. As shown in Figures S1A–D in Supplementary Material available online at http://dx.doi.org/10.1155/2015/916352, gene *ABHD10* that did not show significant difference in gene expression and coefficient curves of the ODE demonstrated strong difference in the unit-step response.

Figures S2A–D plotted the average expression curves, unit-step response curves, and the coefficient curves of the ODE of gene *BTS2*, respectively. Gene *BTS2* was differentially expressed but did not show significant difference in coefficient of the ODE between tumor and normal samples. Gene *BTS2* was identified to have significant difference in the unit-step response. The pattern of difference in the unit-step response may mainly due to rapid changes of gene expression in the region close to genomic position 20. From the literature we found that *BTS2* was associated with a number of cancers [18, 19].

4. Discussion

Dominant methods in literature for RNA-seq data analysis use a single valued summary statistic to represent expression level of a gene. However, a single number oversimplifies complex expression variation pattern across a gene and ignores information on alternative splicing and isoform and expression level variation at the genomic position level. To extract biologically useful expression variation signals across gene from RNA-seq data is a challenge, but important task. To meet this challenge, we have proposed using the ODE for modeling the RNA-seq data and addressed several essential issues for application of the ODE model to RNA-seq data analysis.

The first issue is how to use the ODE for modeling the RNA-seq data. We considered the number of reads or expression level at each position as a function of the genomic position and viewed the transcription process as a stochastic process of transcription along the gene. Borrowing dynamic theory from engineering, we have used the second order ODE to model the expression function of the gene measured by RNA-seq. We have employed differential principal analysis to develop statistical methods for estimation of location-varying coefficients of the ODE. We observed that the second order ODE almost has as good accuracy to predict gene expression as the third order ODE. But the third order ODE requires one more degree to describe the model. Therefore, the second order ODE is good enough to model gene expression.

The second issue is the precision of the ODE to model the RNA-seq data. We randomly split normal and cancer samples into five groups. From the estimation of parameters in the training samples, we solved the ODE with estimated coefficient functions to predict the expression curves of test samples. We have showed that the accuracy of the prediction by the second order ODE was very high and the root mean square prediction errors were quite small.

The third issue is how to extract useful regulatory signals from the RNA-seq data confound with measurement errors and sequencing technology variation. Since the second order ODE can model RNA-seq data very well, the location-coefficient functions of the ODE may well characterize the features of the regulatory process and measure the impact of the gene expression on the function of the cells and tissues. We have demonstrated that using location-coefficient functions of the second order ODE as features we have accurately classified the tumor and normal samples.

The fourth issue is to explore the applications of the ODE for RNA-seq data analysis. We have showed that the ODE can be used as a powerful tool to study the response of the gene transcription to the perturbation of environments. We have identified a number of cancer associated genes which showed significant difference in the response of the gene transcription between tumor and normal tissues but were not differentially expressed.

To our knowledge, this is the first time to use the ODE for modeling the RNA-seq data and investigation of gene transcription process. Our results were very preliminary. The samples were used to validate the accuracy of the ODE model to fit the real RNA-seq data. Large-scale validation and experiments for evaluating the model precision are urgently needed. Although the response analysis of dynamic model for the transcription process can help us to study how the external signals affect the gene expression variation across the gene, the mechanism of the gene transcription variation under the perturbation of external signals is largely unknown. The experiments for validation of the results of the response analysis of the dynamic models need to be performed. We lack consensus methods for RNA-seq data analysis. We are facing great challenges in developing innovative approaches and general framework for RNA-seq data analysis.

5. Conclusions

In conclusion, this study proposes the second order ODE for modeling RNA-seq data. We have demonstrated that the estimated ODE can accurately predict the gene expression level across the gene. We have showed that the location-dependent coefficients of the ODE effectively extract regulatory signals from the RNA-seq confounded with the measurement errors and sequencing technology variation and capture the inherent features of the transcription process. The results have showed that using coefficients of the ODE as features we can reach very high accuracy for classifying tumor and normal samples. Finally, we have demonstrated that using transient response analysis of dynamic system we identified 31 genes with significant differential response behavior between tumor and normal samples related to cancer.

Conflict of Interests

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

Acknowledgments

The project described was supported by Grants 1R01AR057120-01 and 1R01HL106034-01 from the National Institutes of Health and NHLBI. The authors wish to thank TCGA working group for providing RNA-seq data.

The authors wish to acknowledge the contributions of the research institutions, study investigators, field staff, and study participants in creating the TCGA datasets for biomedical research.

References

[1] H. Xiong, J. B. Brown, N. Boley, P. J. Bickel, and H. Huang, "DE-FPCA: testing gene differential expression and exon usage through functional principal component analysis," in *Statistical Analysis of Next Generation Sequencing Data*, pp. 129–143, Springer, New York, NY, USA, 2014.

[2] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 2011.

[3] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[4] F. Rapaport, R. Khanin, Y. Liang et al., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biology*, vol. 14, no. 9, article R95, 2013.

[5] C. Suo, S. Calza, A. Salim, and Y. Pawitan, "Joint estimation of isoform expression and isoform-specific read distribution using multi-sample RNA-Seq data," *Bioinformatics*, vol. 30, no. 4, Article ID btt704, pp. 506–513, 2014.

[6] Z. Sun and Y. Zhu, "Systematic comparison of RNA-Seq normalization methods using measurement error models," *Bioinformatics*, vol. 28, no. 20, pp. 2584–2591, 2012.

[7] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Research*, vol. 38, no. 12, article e131, 2010.

[8] K. D. Hansen, Z. Wu, R. A. Irizarry, and J. T. Leek, "Sequencing technology does not eliminate biological variability," *Nature Biotechnology*, vol. 29, no. 7, pp. 572–573, 2011.

[9] K. Ogata, *System Dynamics*, Prentice Hall, 3rd edition, 1997.

[10] L. Luo, E. Boerwinkle, and M. Xiong, "Association studies for next-generation sequencing," *Genome Research*, vol. 21, no. 7, pp. 1099–1108, 2011.

[11] J. O. Ramsay, *Functional Data Analysis*, Wiley Online Library, 2006.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, 2nd edition, 2009.

[13] K. Koh, S. J. Kim, and S. P. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[15] K. Soetaert, J. Cash, and F. Mazzia, *Solving Differential Equations in R*, Use R!, Springer, 2012.

[16] Y.-H. Liu, C.-Y. Lin, W.-C. Lin, S.-W. Tang, M.-K. Lai, and J.-Y. Lin, "Up-regulation of vascular endothelial growth factor-D expression in clear cell renal cell carcinoma by CD74: a critical role in cancer cell tumorigenesis," *The Journal of Immunology*, vol. 181, no. 9, pp. 6584–6594, 2008.

[17] S.-Q. Ji, X.-L. Su, W.-L. Cheng, H.-J. Zhang, Y.-Q. Zhao, and Z.-X. Han, "Down-regulation of CD74 inhibits growth and invasion in clear cell renal cell carcinoma through HIF-1α pathway," *Urologic Oncology: Seminars and Original Investigations*, vol. 32, no. 2, pp. 153–161, 2014.

[18] K. H. Fang, H. K. Kao, L. M. Chi et al., "Overexpression of BST2 is associated with nodal metastasis and poorer prognosis in oral cavity cancer," *The Laryngoscope*, vol. 124, no. 9, pp. E354–E360, 2014.

[19] A. Sayeed, G. Luciani-Torres, Z. Meng, J. L. Bennington, D. H. Moore, and S. H. Dairkee, "Aberrant regulation of the BST2 (Tetherin) promoter enhances cell proliferation and apoptosis evasion in high grade breast cancer cells," *PLoS ONE*, vol. 8, no. 6, Article ID e67191, 2013.