

## Methodology Report

# Time Series Expression Analyses Using RNA-seq: A Statistical Approach

Sunghye Oh,<sup>1</sup> Seongho Song,<sup>2</sup> Gregory Grabowski,<sup>1</sup> Hongyu Zhao,<sup>3</sup> and James P. Noonan<sup>3</sup>

<sup>1</sup> Department of Pediatrics, Children's Hospital Medical Center, Cincinnati, OH 45229-3039, USA

<sup>2</sup> Department of Mathematical Science, University of Cincinnati, OH 45221-0025, USA

<sup>3</sup> Department of Genetics, Yale School of Medicine, New Haven, CT 06520-8005, USA

Correspondence should be addressed to Sunghye Oh; [sunghye.oh@cchmc.org](mailto:sunghye.oh@cchmc.org) and James P. Noonan; [james.noonan@yale.edu](mailto:james.noonan@yale.edu)

Received 6 October 2012; Revised 10 January 2013; Accepted 15 January 2013

Academic Editor: Tun-Wen Pai

Copyright © 2013 Sunghye Oh et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

RNA-seq is becoming the *de facto* standard approach for transcriptome analysis with ever-reducing cost. It has considerable advantages over conventional technologies (microarrays) because it allows for direct identification and quantification of transcripts. Many time series RNA-seq datasets have been collected to study the dynamic regulations of transcripts. However, statistically rigorous and computationally efficient methods are needed to explore the time-dependent changes of gene expression in biological systems. These methods should explicitly account for the dependencies of expression patterns across time points. Here, we discuss several methods that can be applied to model timecourse RNA-seq data, including statistical evolutionary trajectory index (SETI), autoregressive time-lagged regression (AR(1)), and hidden Markov model (HMM) approaches. We use three real datasets and simulation studies to demonstrate the utility of these dynamic methods in temporal analysis.

## 1. Introduction

RNA-sequencing (RNA-seq) has fundamentally become the choice of studies of transcriptome [1–6]. From the conventional technologies in microarray and beginning of digital sequencing SAGE [7], a significant hurdle in the analysis of the transcriptome arises from insufficient samples, specifically, in identification of the temporal patterns of gene expression measured at a series of discrete time points. Several data-mining techniques and statistical methodologies have proven to be useful to search temporal gene expression patterns in microarrays [3, 8–34]. Some people have already started to adapt the way we applied in microarrays for RNA-seq data. The main drawback, however, is the loss of discreteness property of read count on transcriptional level, albeit there are no additional advantages in analytical aspects on counts. Given experimental design with sufficient replicate, time points, and sequencing depth [4, 5, 34], attempts to RNA-seq specific methodologies to preserve the elegant count property in time course will contribute to development and application in this area ahead. The last four years witnessed the astonishing publication of statistical methodology studies to identify

differential expression between two or amongst multiple groups. Nonetheless most analysis tools remain tied to a static model approach without respect to time, albeit the incisive ultrahigh-throughput sequencing data now provides time series gene expression profile. As the first step towards understanding temporal dynamics in RNA-seq data, temporal analyses often rely on the simple pairwise comparisons [35–47] to infer differentially expressed genes/isoforms at a specific time point versus a reference time point. Differential expression results are then combined to characterize the dynamics over time. Commonly used microarray data analysis methods, such as limma [40], log linear models [39], and ANOVA [26], after variance-stabilizing transformation have also been used for temporal data analysis in RNA-seq as another alternative. However, the very few replications for such data limit the power of these methods. Statistical inference from such high dimensional data structure with the large number of variables and very few observations has presented substantial challenge. More importantly, the pairwise approaches fail to account for the strong temporal dependencies; indeed, higher correlation between neighboring time points is clearly revealed in published gene expression profiles [48] and our

real data applications (see Figure 2). Therefore, these pairwise approaches are suboptimal without explicitly modeling the expression dynamics over time nor can the time points that contribute most to time evolution trajectory pattern of a gene's expression be identified. More descriptive methods, such as clustering methods, have also been applied to identify coexpressed gene sets using RNA-seq data [49–51]. Such unsupervised clustering methods implicitly assume that data collected at different time points are independent, ignoring the sequential structure in time series data. It is apparent that potentially useful information on gene regulation and dynamics may be lost with these suboptimal methods, and there is a need to develop statistical methods that can appropriately model and analyze RNA-seq data. We discuss several methods that explicitly model the time-dependent nature of the time series data in this paper. We describe the identification of temporal differential expression (TDE) analyses as well as the ranking of genes to show temporal trajectories with statistical significance. We also discuss the application of time-lagged autoregressive AR(1) models to identify TDE genes as well as hidden Markov models (HMM) to classify different expression patterns by posterior probabilities of latent states. These methods can be applied to study complex factorial designs that interrogate multiple biological conditions simultaneously where multiple time points are studied under two or more biological conditions. Multivariate approaches are presented to identify temporal patterns in coexpressed gene groups and quantify coupled relationship of two distinct trajectories. Here we report an in-depth analysis of temporal patterns based on nonparametric and Bayesian approaches that incorporate the context of inherent time dependence of gene expression *per se*. When these methods are applied for published real datasets, both static and dynamic methods performed well for most temporal genes; however, dynamic methods had particularly a slight edge at low and moderate expression levels. That may be particularly advantageous for years to come for application to data with relatively low signals such as depression and aging data, which on expression compared to tumors in disease tissues.

## 2. Statistical Methods

**2.1. Time Series Data Structure.** Suppose that a gene expression profile matrix contains  $i = \{1, \dots, G\}$  genes and  $j = \{1, \dots, m\}$ ,  $m$  different time stages. The  $i$ th gene expression profile vector,  $Y_i = [y_i(t_1), \dots, y_i(t_m)]^t$ , corresponds to a sequential vector of time points and biological replicates within a time point, namely, where  $y_i(T = t_j) = [y_{it,L=1}, \dots, y_{it,L=l}]$  is a vector composed of intraexpression measurements by  $L = \ell$  biological replicates at time point  $T = t_j$ . We consider a sequence of observations on gene expression profile dataset, made at  $m$  different time points; accordingly  $m$  dimensional gene expression vector of gene  $i$  with observed read counts over time is used hereafter.  $y_{ij} = [y_{ij1}, y_{ij2}, \dots, y_{ijc}]^t$  is  $C = c$  dimensional gene expression vector of gene  $i$ , time point  $j$ . The expression profile is a factorial time course experiment and the vector  $y_{ij}$  represents the intraexpression profile of  $c$  biological condition within a

time point.  $y_{ijc} = [y_{ijc1}, y_{ijc2}, \dots, y_{ijcl}]^t$  is an  $l$  dimensional gene expression vector of gene  $i$ , time point  $j$ ,  $c$  biological condition, and  $l$  different biological individual replicates. If there are not any treated biological conditions, the gene expression time series is simplified in  $y_{ijl}$ .

**2.2. Statistical Evolutionary Trajectory Index (SETI).** Existing static methods for testing significance of TDE genes in time series RNA-seq data do not consider temporal stochastic ordering dependency property in time, which differs from a typical gene expression profile data, and all static methods assume samples that are distributed independently and are not related to each other instead. However, it is well known that the considerable genes in gene expression profiles related to many developmental biological processes or disease progression are temporally differentially expressed and current expression level is affected by previous one by inherent Markovian property in time series. In the settings of large numbers of variables and with few observation, distribution-free or Bayesian approaches by using useful prior information are more suitable in RNA-seq. To circumvent the limitations and cope with a variety of particular patterns in time course, we present a statistical framework that enables more precise temporal expression profiling by incorporating autocorrelation measurement to determine relationship between consecutive expression profiles. Residuals in one period (or time point) are correlated with those in previous periods (or time point) and ranking individual SETI based on nonparametric regression fit as a gene-by-gene approach. As above, the gene expression level  $y_{ij}$  at  $I = i$ th gene,  $J = j$ th time,  $C = \{1, \dots, C\}$  biological condition, and  $L = \{1, \dots, L\}$  replicates is fitted by smooth spline regression. The autocorrelations of the residuals are computed by the sliding of all possible cases over the original time series, which are referred to as a trajectory index for given gene. The unbiased estimate of the autocorrelation for each gene is

$$\hat{ACR}_{res,i}(k) = \frac{1}{(m-k)\sigma^2} \sum_{j=1}^{m-k} [Y_{ij} - \hat{Y}_i] [Y_{ij+k} - \hat{Y}_i] \quad (1)$$

for any positive integer  $k < G$ .  $\{y_1, y_2, \dots, y_G\}$  is a vector to be contained by  $G$ -length observations of expression measurement.  $P$  values for assessing statistical significance are calculated using a permutation test ( $N = 10,000$ ), assuming the absence of temporal differential expression. The confidence interval and trimmed mean of trajectory index are derived by bootstrapping analysis ( $B = 100$ ). The method is based on computing autocorrelations, that is, cross-correlation of gene expression profile across time points to represent temporal pattern. It is applied in a variety of different types of RNA-seq time series data including factorial time course experiments.

**2.3. Autoregressive Time-Lagged AR(1) Model.** We propose to use an autoregressive time-lagged AR(1) model for the identification of temporal and differential gene expression. Hay and Pettitt [54] demonstrated first-order time lag for an application to the control of an infectious disease with

count data over time in which the time series observations are examined to identify significant associations with explanatory variables and counts, the incidence of an infectious disease ESBL-producing *Klebsiella pneumoniae* in an Australian hospital, and the explanatory variable is the number of grams of antibiotic third-generation cephalosporins used over that time period. In order to essentially propose a universal dynamic method with AR(1) model in RNA-seq, we consider models to allow flexibility without covariates in lieu of taking their initial approaches. The details of our AR(1) model for read count gene expression profile over time as a gene-by-gene TDE identification are discussed in the following with mathematical notations. Bayesian framework is defined by  $(y_{ij} | \mu_{ij}, i = 1, \dots, n \text{ and } j = 1, \dots, m)$  to be independently distributed as Poisson model. We employ their model for RNA-seq read count expression data.

**2.3.1. Poisson Model in AR(1).** From the time series data structure (Section 2.1), we have  $m$  time points,  $c$  biological conditions, and  $l$  replications. Both maize and zebrafish data with single measurements within a time point are applied in this method.

Consider

$$y_{ij} \sim \text{POI}(\mu_{ij}), \quad \text{where } i = 1, \dots, n, \quad j = 1, \dots, m,$$

$$\log(\mu_{ij}) = w_{ij} + \beta_i,$$

$$w_{i1} = \frac{u_{i1}}{\sqrt{1 - \varphi_i^2}}, \tag{2}$$

$$w_{ij} = \varphi_i w_{ij-1} + u_{ij}, \quad j > 2.$$

And equivalently,

$$\log(\mu_{ij}) = w_{ij} + \beta_i,$$

$$w_{i1} \sim \text{Normal}\left(0, \frac{\sigma^2}{(1 - \varphi_i^2)}\right), \tag{3}$$

$$w_{ij} | w_{i1, \dots, j-1} \sim \text{Normal}(\varphi_i w_{ij-1}, \sigma^2), \quad j > 2.$$

To identify altered gene expression across time series, for each gene, the AR(1) model is applied and inference of  $\beta$  is obtained from noninformative priors and time series random effects for sequential expression profile are assumed. This autoregressive model was originally carried out for longitudinal large-scale historical repeated-measurement data. In our study, using the modified assumptions, RNA-seq time series with short time period (4~8 time points) and single observations as gene-by-gene approach are applied to compare the performance of AR(1) model to static methods in identification of differential expression. The posterior probabilities of parameters in the model are estimated through MCMC simulations with  $N = 6,000$  iteration and 1,000 burn-in. We provide detailed notations and equations for three dynamic approaches in Supplementary data available online at <http://dx.doi.org/10.1155/2013/203681>. In the results, we are most interested in autocoefficients to represent time series

sequential random effects in the model and we implemented a classification between TDE (temporally differential expression over time) and EE (equally expression over time) set of genes. Similar to statistical differential expression testing for each gene in a classical approach, our implementation of testing in AR(1) model is given by a Bayesian interval estimate, 95% credible interval:

$$H_0 : \text{if } \varphi_i = 0, \text{ EE}, \quad H_1 : \text{otherwise, TDE}, \tag{4}$$

where we consider that gene  $i$  is temporally differentially expressed (TDE) if the 95% credible interval of  $\varphi_i$  does not include 0; otherwise it is considered to be equally expressed (EE). Also we obtain the tail probability of  $(\varphi_i | y)$  of gene  $i$ , that is,  $p(\varphi_i > 0 | y)$  or  $p(\varphi_i < 0 | y)$  for  $i = 1, \dots, n$  using MCMC. It indicates the significance of differential expression for each gene.

**2.3.2. Negative Binomial in AR(1).** A more compelling methodological goal is to infer temporal dynamics when we have replicates within a time point and it is straightforward to establish a negative binomial model with AR(1):

$$y_{ij} \sim \text{NBC}(k, \mu_{ij}), \quad \text{where } i = 1, \dots, n,$$

$$j = 1, \dots, m, \quad \log(\mu_{ij}) = \omega_{ij} + \beta_j. \tag{5}$$

Other parts of the model remain identical as in (3). Here,  $y \sim \text{NBC}(k, \mu)$  means that  $y$  has its probability function as follows:

$$p(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k$$

$$\times \left(1 - \frac{k}{\mu+k}\right)^y, \quad y = 0, 1, 2, \dots \tag{6}$$

This negative binomial distribution has its mean  $E(y) = \mu$  and its variance  $\mu + \mu^2/k$ . The parameter  $k^{-1}$  is called the dispersion parameter.

**2.4. Hidden Markov Model (HMM).** We consider a Bayesian HMM to analyze factorial time course RNA-seq data. Our model follows the seminal work of Yuan and Kendziorski [48] that characterizes all possible temporally differential expression patterns in time series microarray data with two or more biological conditions. Although this early study was encouraging, the HMM was restricted to represent timing differences between biological conditions with binary EE/DE or multiple cases of latent hidden states depending on the number of given conditions at each time point. The extent of temporal changes was not obvious in significantly differentiating between one time point and the next. Taking a HMM approach, we seek SETI and multivariate coupled relationships among distinct trajectories into HMMs in each condition to investigate biological evolutionary trajectory that can be applied to a comprehensive set of RNA-seq time series data to make probabilistic predictions of temporal patterns for how differential expression will occur under different biological conditions. Also, count specific underlying

distributions for RNA-seq time series data are used. First, we introduce a mechanism to use the inference of temporally differentially expressed genes in time series RNA-seq gene expression profiles with multiple biological conditions at a given time point. This was achieved by incorporating GP and NBD with corresponding prior information into the HMM for each gene, allowing samples having either multiple replicates or single observations. We investigate properties of the HMM technique such as how it benefits by incorporating hidden variables when making the predictions of temporal patterns of differential expression for given different biological conditions and how the number of chosen latent variables varies with conditions within a stage over a time period. As per Section 2.1, we present how to express hidden states in the given models with subindices composed of  $T$  time points,  $C$  different biological conditions (e.g., drug treatments or tissues), and  $L$  replicates. As RNA-seq experiments generally have small sample sizes, the identification of statistically significant temporally differentially expressed (DE) genes may have limited power. Also, some studies stress the importance of replication in microarray studies, which have inherent variability [4, 5, 33, 34, 55] regardless of how well constructed DE methods are applied. Thus, without replicates, no statistical significance tests are reliable and powerful on detection of TDE. With the reduction in sequencing costs, well-designed balanced RNA-seq experiments with proper sample sizes and time points will facilitate the use of temporal dynamic methods, including AR(1) model. Here HMM is used with samples and 4 biological conditions (different tissues). Consider that the gene expression dataset  $(y_{ijcl})$  has  $I =$  genes,  $J =$  time points,  $C =$  conditions, and  $L =$  replicates. This algorithm has the Markovian assumption that the expression level at the current time only depends on that at the most recent time. We use hidden states to represent a change in expression levels between different biological conditions. Thus, this framework allows us to detect TDE genes and to facilitate the calculation of the posterior probabilities of all possible TDE patterns. For instance, with three time points, this method can estimate the posterior probability of pattern EE-DE-EE, where EE stands for equally expressed and DE for differentially expressed, respectively. Namely, the main interest is to identify the relationship among the  $C$  class latent mean values of expression level for each gene  $g$  at each time point  $T = t$  denoted by  $\mu_{gt1}, \mu_{gt2}, \dots, \mu_{gtC}$ . Hereby, the primary goal of HMM in time course experiment with multiple different conditions is to infer all potential relationships from different conditions; for simplest case with two biological conditions, it is binary outcome with EE/DE, and for complicated experimental design with more than two biological conditions, suppose that biological conditions correspond to different tissues, hereafter tissues A, B, C, and D. Correspondingly, there are 4 expression profiles  $\mu_{gtA}, \mu_{gtB}, \mu_{gtC}$ , and  $\mu_{gtD}$ , and 15 possible expression pattern states include the following:

$$\text{State 1 [1111]} : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$\text{State 2 [1221]} : \mu_1 = \mu_4 \neq \mu_2 = \mu_3$$

$$\text{State 3 [1222]} : \mu_1 \neq \mu_2 = \mu_3 = \mu_4$$

$$\text{State 4 [1121]} : \mu_1 = \mu_2 = \mu_4 \neq \mu_3$$

$$\text{State 5 [1212]} : \mu_1 = \mu_3 \neq \mu_2 = \mu_4$$

...

$$\text{State 14 [1233]} : \mu_1 \neq \mu_2 \neq \mu_3 = \mu_4$$

$$\text{State 15 [1234]} : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

(7)

More generally, the number of all potential patterns as a function of the number of tissues is equal to the Bell exponential number of possible set partitions. Here each state is not observed and needs to be estimated from the data. Therefore, we refer to such states as hidden. For each gene  $g$  at each time point  $T = t$ , we want to estimate the probability of each hidden state  $p(\vec{g}_{gt} = k)$  and then we associate an observation model with each state and eventually also compute the most likely sequential states over time to derive timing differences for a given gene  $g$ . Fitting a hidden Markov model involves estimating the transition probability matrix  $A$ , initial probability distribution  $\pi_0$ , and unobserved hidden state at time  $T = t$ , and estimations are done by EM algorithm as described and implemented in the original paper of HMM. The parametric empirical models (PEM) of GP and NBD sample  $y = (y_1, y_2, \dots, y_N)$  are considered here.

In the GP model, for two biological conditions at each time point and two marginal distributions of hidden states are given the following equations, as shown in Yuan et al., for microarray application. The underlying distributions and joint predictive density (JPD) for discrete count data are incorporated to infer posterior probability distributions:

$$f_{1t}(x_{gt}) = f_{0t}(x_{gt} | \lambda_{gt}) dG_t(\lambda_{gt}) \quad (8)$$

under EE state 1 and

$$f_{2t}(x_{gt}) = \int f_{0t}(x_{gt1, \dots, gt n_1} | \lambda_{gt}) dG_t(\lambda_{gt}) \\ + \int f_{0t}(x_{gt(n_1+1), \dots, gt(n_1+n_2)} | \lambda_{gt2}) dG_t(\lambda_{gt2}) \quad (9)$$

under TDE state 2.

If  $\pi_i$  represents the proportion of TDE genes at time  $t$ , then the mixture type of marginal distribution of the data is given by

$$(1 - \pi_1) f_{1t}(y_{gt}) + \pi_2 f_{2t}(y_{gt}), \quad \text{where } i = 1, \dots, d. \quad (10)$$

And  $f_{0t}(y | \mu_{gt}) = \lambda \exp(-\lambda y) / y!$ ,  $x > 0$ .  $\lambda_t$  follows a conjugate prior with gamma distribution parameters, shape parameter  $\alpha_t$ , and rate parameter  $\beta_t$ . Thus, three parameters  $\theta_t = (\lambda_t, \alpha_t, \beta_t)$  need to be estimated for a given gene. For the GP model, the Markov chain is assumed to be homogeneous and the marginal distribution of  $x_{gt}$  is the finite mixture  $\sum_{i=1}^d \pi_i f_{it}$ . We assume one-step first-order correlation time series structure so that HMM contains with

Poisson distributed state-dependent distribution. The goal of this algorithm is to identify a certain set of genes that are TDE in a combination of time series and four different biological conditions, for example, distinct tissue types. To address the utility of HMMs proposed in time course RNA-seq experiments with multiple different tissues, we exploit a parametric hierarchical empirical Bayes model with GP (data w/o replication) and NBD (data w/replications) with beta-prior as a well-modified Bayesian approach [42, 56, 57]. The Newton et al. [57] approach identifies differentially expressed genes for microarray experiment framework in multiple biological conditions at a static time point and similarly Hardcastle and Kelly [42] identify differentially expressed genes either for pairwise comparisons or for multiple group comparisons in an RNA-seq experiment framework at a static time point. For microarray data, Yuan and Kendziorski [48] proposed a HMM for a dynamic time course experiment with multiple conditions Gamma-Gamma (GG) and Log Normal Normal (LNN) to identify genes of interest whose temporal profiles are different across two or more biological conditions. Here, we adapted and extended that approach to a general RNA-seq framework with GP and NBD models as more flexible models. The earlier studies are limited to detect temporal patterns other than ranking/ordering temporal dynamic specific genes during developmental stages, which biologists are more interested in examining. We assume two common underlying distributions for RNA-seq read count. In reality, violation of GP assumptions is very common and in order to account for overdispersion. Alternatively, NBD is applied with a beta-prior. The above inference method provides for continuous trajectory regression involved with timing evolution features to rank temporal genes statistically for a given pattern, as well as such genes' temporal differential expression patterns among conditions. In addition, we examined multivariate identification of temporal expression using the following several metrics.

### 2.5. Coupled Multivariate Identification of SETIs

**2.5.1. Granger Causality.** The concept of Granger causality between two distinct SETIs assumes that the data at the current time point affect the data at the succeeding time point [58]. To determine Granger causality for each pair of trajectories, we employ standard *F*-statistics to test if the residual values derived from the fitting smoother for gene A are incorporated into the equation for another gene B. If all the coefficients for the measurements of gene B are zero under the null hypothesis, then there is no statistically significant Granger causality between the trajectories for genes A and B.

**2.5.2. Cotrajectory with Glass-d-Score.** Similarly, each pair of two trajectories, which correspond to two gene expression levels, is explored by another dependency metric score and detailed notations are described in the following, when there is a given pair of two gene expression profiles:

$$(g_i, g_j) d_k^{ij} = \frac{r_k^{ij} - \bar{r}_k^i}{\sigma_{r_k^i}}, \quad (11)$$

where  $r_k^{ij}$  is the correlation coefficient between the expression profiles of  $(i, j)$  among all possible pairs. The null distribution was assumed to have  $\bar{r}_k^i (\sigma_{r_k^i})$  the mean and standard deviation of correlation coefficient between gene  $i$  and all other genes, respectively.

**2.5.3. Correlation Approach.** As proposed in Ma et al. and Barker et al. we propose a biologically motivated approach to measure the relationship between two different genes based on their temporal expression profiles in RNA-seq. Ma et al. proposed to consider lagged coexpression analysis to capture the scenario that there is a delayed response of gene B to gene A so that the profile of gene B is correlated with the time delayed profile of gene A.

**2.6. Pairwise Methods.** In this section, we describe the pairwise methods that we consider in our comparisons with the methods discussed above that can explicitly model the time dependencies nature in the data. For comparisons with our dynamic methods, we examined several popular static methods, including Fisher's exact test for simple two sample comparisons and log linear model for multigroup comparison, which can also be applied for RNA-seq time series data in temporal analysis as intuitive but limited.

DE analyses: we first employed pairwise condition comparison methods in digital measures at a given static status without respect to time. It is no surprise to take a union set of all possible pairwise comparisons using these static techniques to identify temporal dynamics in relatively small experiments, where single sample for each time point and very few number of time points are contained in experimental design.

- (i) Fisher's exact test: from Table 1, the 2-sided *P* value for TDE of each gene is computed with (12) [39]:

$$\Pr(g_{+1,g} = g \mid g_{+1}, g_{+2}, \dots, g_{.g}) = \frac{\binom{g_{+1}}{g} \binom{g_{+2}}{g_{.g}-g}}{\binom{g_{.}}{g}}. \quad (12)$$

- (ii) Audic-Claverie statistics.

The Audic-Claverie statistics [59] are based on a distribution  $p(y \mid x)$  over read counts  $y$  in one sample in one given group informed by the read counts  $x$  under the null hypothesis that the read counts are generated identically and independently from an unknown Poisson distribution.  $p(y \mid x)$  is computed by infinite mixture of all possible Poisson distributions with mixing proportions equal to the posteriors under the flat prior over  $\lambda$ . When the two libraries in a given Solexa/Illumina RNA-seq experiment are of the same size,

$$p(y \mid x) = \frac{1}{2^{x+y+1}} \frac{(x+y)!}{x!y!} = \frac{1}{2^{x+y+1}} \binom{x+y}{x}. \quad (13)$$

These are Audic-Claverie statistics [59] for given read counts  $x$  and  $y$ .

Pooling methods: as with ANOVA in microarray, log linear model and linear models for microarray data (LIMMA),

TABLE 1:  $2 \times 2$  contingency table.

Tags in gene	Reads from sample of type		
	Group A	Group B	
Gene g	$\mathcal{G}_{+1,A}$	$\mathcal{G}_{+2,A}$	$\mathcal{G}_{.,A}$
Not gene g	$\mathcal{G}_{+1} - \mathcal{G}_{+1,A}$	$\mathcal{G}_{+2} - \mathcal{G}_{+2,A}$	$\mathcal{G}_{.,B}$
Total	$\mathcal{G}_{+1}$	$\mathcal{G}_{+2}$	$\mathcal{G}_{..}$

after variance-stabilizing transformation to allow multigroup and multifactor comparisons, can be applied by including a time variable as the main factor in the model [40].

- (i) Log linear model with the Poisson link function (or negative binomial when replicates are available) and likelihood ratio test model. In the model, the time factor, biological condition factors, and their interaction terms are included.
- (ii) LIMMA (linear model for microarray) with  $F$ -statistics under the linear model setting implemented in R package is also applied for time series RNA-seq read count data after variance stabilizing transformation.

Although such static algorithms have demonstrated a successful identification of temporally expressed genes in some degree in the past four years and our study, any temporal dynamic analysis false discovery results in static methods can be introduced due to violation of Markovian assumptions frequently revealed in time series expression profile. As the cost to sequencing continues to decline, there is urgent need for more sophisticated statistical methodologies of power in the identification of temporal expression or for use of characterization of temporal dynamics to assess isoform diversity within a gene level in a future investigation of time series RNA-seq. Ideally, it is very critical to appropriately have a good model to represent observed data since interpretation of a model that does not contain valuable information is useless. For this important purpose, our dynamic methods are compared to these static methods by evaluating the overlap in the number of differentially expressed genes in real data sets.

### 3. RNA-seq Time Series Data

**3.1. Three Different Types of Time Series.** There are mainly two types of time series in RNA-seq. The first is factorial time series data that include at least two biological conditions to be compared in a given time point and have multiple developmental patterns over time as the number of conditions. The second type of time series has a single condition and corresponding developmental stage. In the third type of time series, there are subsequently two additional subtypes, circadian rhythmic data and cell cycle data. In this study, we formulate the statistical framework of identification of temporal changes in RNA-seq time series for first two types of data and the periodic data-sets are reviewed in “another review manuscript” with discrete Fourier transformation and other methods in a separation in depth.

### 3.2. RNA-seq Real Time Series Datasets

**3.2.1. Factorial Time Course Experiment: A Sheep Model for Delayed Bone Healing.** We consider this published RNA-seq time series data from a sheep model for delayed bone healing. In Jager et al., surgery was conducted as described in [52, 53] and the newly generated tissues were harvested at different days, 7, 11, 14, and 21 after surgery. For each time point, there are 6 biological replications for both groups except one time point, for day 21 (group I,  $n = 5$ , group II,  $n = 6$ ), where two groups are defined by standard healing system and delayed healing system. Thus, the authors considered two treatments: standard healing system and treatment with unstable external fixator leading to delayed bone healing. While the standard bone healing system was investigated in a 3 mm tibial osteotomy model stabilized with a medially mounted rigid external fixator, delayed healing was investigated in a 3 mm tibial osteotomy model stabilized with a medially mounted rotationally unstable external fixator. For each treatment, RNA-seq data were collected at 4 time points: 7, 11, 14, and 21 days, with 5-6 individuals’ DNA samples pooled together at each time point. In their differential expression, they used the pooled samples from 5-6 lanes of animal samples at one time point and Audic-Claverie statistics were performed using 4 samples over 4 time points by taking a union set of all possible pairwise comparisons using static methods. We reanalyzed their sheep animal time series data using three dynamic methods to identify TDE genes.

**3.2.2. Single Transient Time Course Experiment-I.** We applied two single biological condition time series data which are interested in exploring developmental transient patterns during a time period rather than timing difference patterns incorporated with multiple conditions at a time as Section 3.2.1 example. Maize leaf transcriptome with four different developmental zones containing two replicates in each time point [50] was employed. This is one representative for time course experiment with single transient expression profile. Tissues were collected from leaf 3 at 9 days after planting 3 hours into the  $L$  period from four segments: (1) basal (1 cm above the leaf three ligule), (2) transitional (1 cm below the leaf two ligule), (3) maturing (4 cm above the leaf two ligule), and (4) mature (1 cm below the leaf three tip). Thus, maize leaf data with different developmental stages are generated from mRNA isolated from four developmental zones: basal zone, transitional zone, maturing zone, and mature zone. In the differential expression analysis, they simply applied chi-squared static method and  $K$ -means clustering method that both do not take into account time dependency, but all samples are assumed to be independent. This maize leaf time series data are reanalyzed with proposed methods in this study.

**3.2.3. Single Transient Time Course Experiment-II.** This is a time series experimental design to be composed of eight stages during early zebrafish development, embryogenesis [51]. In their study, wild-type zebrafish embryos (TLAB) were

TABLE 2: Statistical evolutionary trajectory index (SETI) of the top candidate genes where FDR is controlled at less than 0.05 in a sheep model data. The gene expression level is fitted on smooth spline function, autocorrelation of residuals is measured, and corresponding statistical significance is tested. In addition, trimmed mean of bootstrap and 95 percent of confidence interval are also provided in the table.

Top candidate genes	SETI (trimmed mean of bootstrap)	Bias of bootstrap	95% CI of SETI	P	FDR
A5D9H5	1.23 (1.23)	0.12	[1.02, 1.43]	0	0
A6QQB6	1.23 (1.27)	0.09	[1.00, 1.46]	0	0
A0JN96	1.23 (1.23)	0.09	[1.03, 1.43]	0	0
DUFFY	1.23 (1.23)	0.13	[1.02, 1.43]	0	0
A6QP68	1.23 (1.23)	0.10	[1.03, 1.42]	0	0
gi 11992112	1.23 (1.23)	0.10	[1.05, 1.40]	0	0

staged according to standard procedures and about 1,000 embryos were collected per stage (two to four cells, 1,000 cells, dome, shield, bud, 28 hpf, 48 hpf, and 120 hpf) within a tight time window of ~10 min. Their collection of embryos was ensured that all embryos were at the same developmental stage. The identification of long noncoding RNAs (lncRNAs) expressed during zebrafish embryogenesis was explored to assess a diversity of transcripts that are structurally similar to, but noncoding, mRNAs. The analyses of RNA-seq time series expression profiles focused on the identification of temporal dynamics of lncRNAs using the Cuffdiff method in its time series mode with upper quantile normalization, which is also limited to pairwise comparison from previous time point to right next time point. Here, the data reanalyzed the transcriptomic gene expression profile data with 28,520 annotated protein coding genes. To consider the possibility of similarities and differences in comparisons between static and dynamic methods for time series RNA-seq data, we systematically compared both methods with these data.

3.3. Results in Differential Expression Analysis on Static and Dynamic Methods. For the sheep data, the authors applied the Audic-Claverie method to the normalized expression values, RPKM, to compare later time points to the reference time point (7 day) in both groups. After all pairwise comparisons, they combined the sets of differentially expressed gene sets with 884 genes detected in total from 24,325 mappable genes. Based on these 884 genes, they performed hierarchical clustering to identify gene clusters. Each cluster was then subject to gene ontology analysis to find significant biological functions. The differential analysis performed in original paper is based on static differential analysis method. We reanalyzed their sheep factorial time course experiment data to identify TDE genes over time through dynamic methods, HMM, SETI, and AR(1) model to account for correlated time-dependency structure. HMM identifies temporal patterns with classification of DE/EE at each time point by posterior probabilities, whereas SETI with statistical significance from permutation resampling

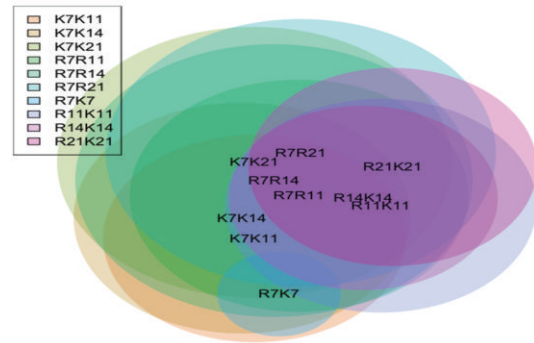


FIGURE 1: Venn diagram for differentially expressed gene sets detected by Fisher’s exact test where the Benjamini-Hochberg FDR is controlled at <0.05. In this figure, as the labels authors used in [52, 53], K represents standard and R represents delayed healing system in a sheep model for two different bone healing systems.

procedures and AR(1) model with gamma Poisson Bayesian assumption on count data are applied within single biological condition, separately. Results obtained by these dynamic methods compared those of static methods, simple pairwise methods, Audic-Claverie statistics and Fisher’s exact test, and pooling static methods, glmFit in edgeR, LIMMA, and log linear model as shown Figure 3. To identify temporal dynamics by assuming correlated data structure, we performed HMM modeling with Poisson-gamma since there were no replicates. AR(1) model and SETI significance tests were also done within each biological condition. Temporally differential expression gene sets detected by these dynamic methods were compared with the results of simple pairwise tests and pooling methods. From the HMM, 646 temporal dynamics of DE calls are identified to represent DE in at least one time point. The HMM model only explores different temporal patterns of DE/EE states and does not rank the genes by statistical significance, but is classifying gene expression profile into a number of temporal patterns by posterior distribution of latent states. Because of this limitation, we

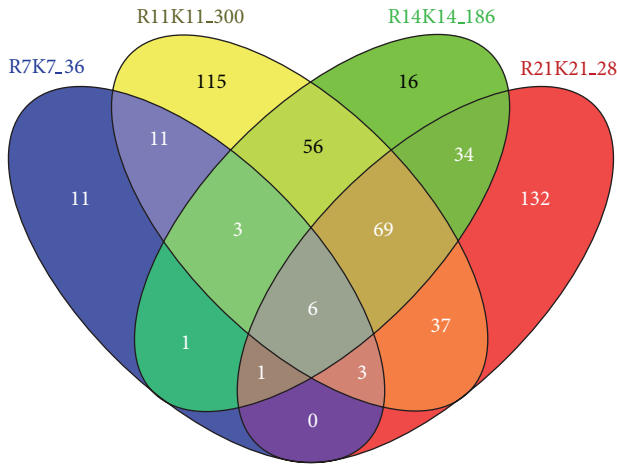


FIGURE 2: Venn diagram of four DE sets having the number of DE genes between two different healing systems detected by Fisher's exact test with FDR 0.05 at each time point ( $t = 7, 11, 14,$  and  $21$  days). Four time points were compared in simple pairwise comparison between two biological conditions, R (delayed healing) versus K (standard healing system). The label of each set depicts the number of DE genes in the specific comparison. The majority of interaction sets of DE genes between two successive time points implies that high proportion in detected differentially expressed genes at current stage tends to be redetected at next stage revealed by inherent time-dependent structure in time series gene expression profile.

employed the SETI and AR(1) models to discover developmental transient patterns in each condition. The trimmed mean time evolution trajectory index is presented for the top three candidate temporal genes in each bone healing system. The 95% confidence interval of bootstrapping and FDR of permutation re-sampling are shown in Figure 4 and Table 2. To determine temporal dynamics and meaningful biological functions, only HMM-specific TDE genes which are not contained in static methods are further explored in gene clustering and biological functional network analysis as shown in Figures 5(a) and 5(b), respectively. In the results, they obviously showed temporally differential expression implying that loss of information to assumption of stochastic time-dependent structure might lead to false discoveries and less power of detection. To discover temporal transient patterns of differential gene expression within each biological condition; healing system, we performed SETI and AR(1) model approaches for each condition, SETI results are given in Figure 4 showing top candidate TDE genes, of which some genes such as *gil119921123* and *B6DXC7* are of low expression levels which we were not able to detect in static methods. In the second data for our study, we have reanalyzed maize leaf transcriptome data to identify TDE genes with static and dynamic methods and compare between two. In their paper, they investigated leaf development gradient in time series gene expression data at successive stages (4 time points: base, tip: basal, transitional, and maturing) and identified a gradient of gene expression from base to tip: basal (23,354) >

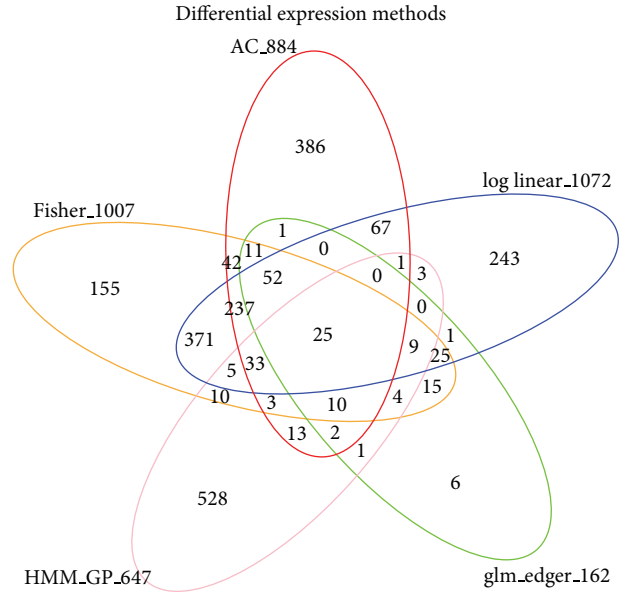


FIGURE 3: Venn diagram of five DE sets for static and dynamic methods in a sheep model data. In static methods, for two simple pairwise methods, Audic-Claverie statistics and Fisher's exact test were performed and both methods take a union set of all possible pairwise comparisons to identify temporally differentially expressed (TDE) genes across time points and two healing systems. As another static approach, pooling methods of samples, log linear model in [39], and generalized linear model fit in edge R in [36] were performed and detected TDE genes by FDR 0.05. In dynamic HMM method, we identify top candidate TDE genes defined at least showing DE pattern from one time point based on posterior probabilities for latent variables (DE/EE between given biological conditions). On the basis of comparison of the number of DE genes identified by each method, patterns of identification of TDE genes are method specific suggesting validation procedures of methods in biological aspects.

transitional (22,663) > maturing (22,036) > mature (21,332) from a total of 25,800 annotated genes. In the differential analysis in times series RNA-seq data, they used the method proposed in Marioni et al. [40] for pairwise analysis. A total of 16,502 genes were found to be differentially expressed in at least one of the comparisons. They then performed K-means clustering and showed eighteen clusters along the four developmental zones (Base, -1 cm, 4 cm, Tip). To compare gene sets detected by our dynamic methods with their gene lists, dynamic methods, SETI, and AR(1) model are applied again in this study and all temporally differentially expressed genes are presented in Supplementary Tables 1 and 3, where filtered gene set to be tested in differential expression has 5273 and 12,322 temporal dynamic transcripts from 42399 transcripts through SETI and AR(1) model, respectively. On the basis of significant temporal expression, we compared dynamic methods to static methods, which were used in the original paper without accounting for correlated data structure type. As the third real data application, to identify temporal dynamics, we have reanalyzed the third data, zebrafish embryonic transcriptome, focusing specifically on



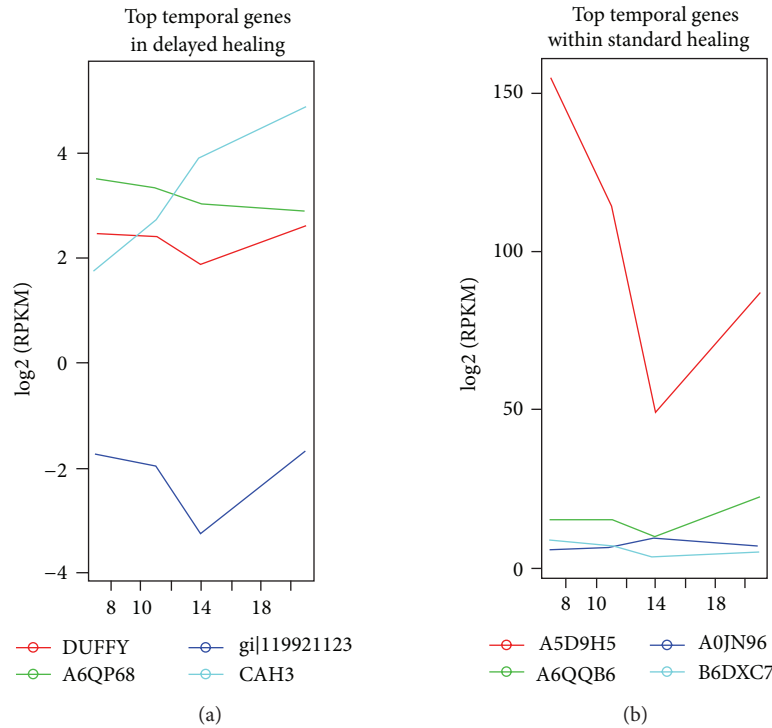


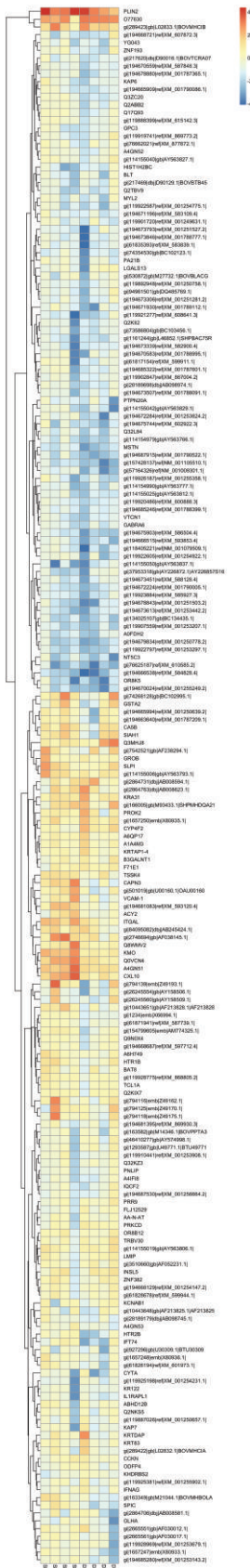
FIGURE 4: Top candidate temporally differentially expressed genes identified by statistical evolutionary trajectory index (SETI) within each healing system in a sheep model data. Each panel depicts temporal patterns between  $\log_2$  (normalized expression levels, RPKM) and four different time points under their expression curves. The distinct colors represent significant individual genes ranked by SETI and FDR 0.05 by resampling procedures.

the identification and characterization of temporally differential expression using statistical evolutionary trajectory index and autoregressive time-lagged AR(1) model. We furthermore implemented both methods to rank temporal genes by statistical significance. As consequence of the resampling-based procedures and posterior probabilities of autocorrelation, it was possible for gene-by-gene approach to order temporal genes by two dynamic methods and identify genes associated with cotemporal dynamics. To investigate such paired temporal dynamics, we examined the relationships between genes using bivariate identification methods. Glass-d score is reported in Supplementary Table 5. Likewise, the statistical evolutionary trajectory index with statistical significance for zebrafish data is given in Supplementary Table 2, where we filtered out genes by coefficient of variation (CV) criteria remaining 12,034 genes. Overall, both methods show more robustness at low and moderate expression levels when compared to existing parametric static methods indicating that our methods achieve relative improvements in test of identification of temporal genes and AR(1) model shows more sensitive TDE calls than SETI resampling procedure in two real data applications. Here, we examined how different results are obtained by dynamic time series methods. For simple pairwise static methods, we employed Audic-Claverie statistics and Fisher's exact test as these two methods have been widely used in previous studies. They showed highly concordant results on other RNA-seq datasets compared to DEGseq, DESeq, edgeR, and baySeq (data not shown). In

differential analysis with simple pairwise methods, we took a union set after all pairwise comparisons across a time period and amongst different biological conditions as these methods only consider two pairwise comparison testing and confirm the results to those of original papers. For pooling static methods, LIMMA, log linear model, and edgeR R package with glmFit are carried out to identify TDE genes. To compare with above static methods, we employed three dynamic methods described in the previous sections. The results are shown in Figures 1 and 3. Figure 2 shows how dependent structure is observed in patterns identified across time points, 36(23), 300(277), and 186(134), of the previous TDE gene set, genes in 64% ~92% percentage are differentially reidentified at the right next time point, implying that there is temporal dependent structure in sheep healing system RNA-seq time series data.

### 3.4. Bivariate Dynamic Analysis for RNA-seq Time Series.

In systems developmental biology where characterization of complexity of various time course data likely leads to address inference of temporal dynamic patterns from transcriptome, we are not often really interested in exactly how only a single gene is temporally differentially expressed at a particular time point or period. This knowledge would neither answer an understanding of how biological networks in temporal dynamics of gene regulation work nor enable predicting any cooperative sets of genes to occur under biological



(a)

FIGURE 5: Continued.

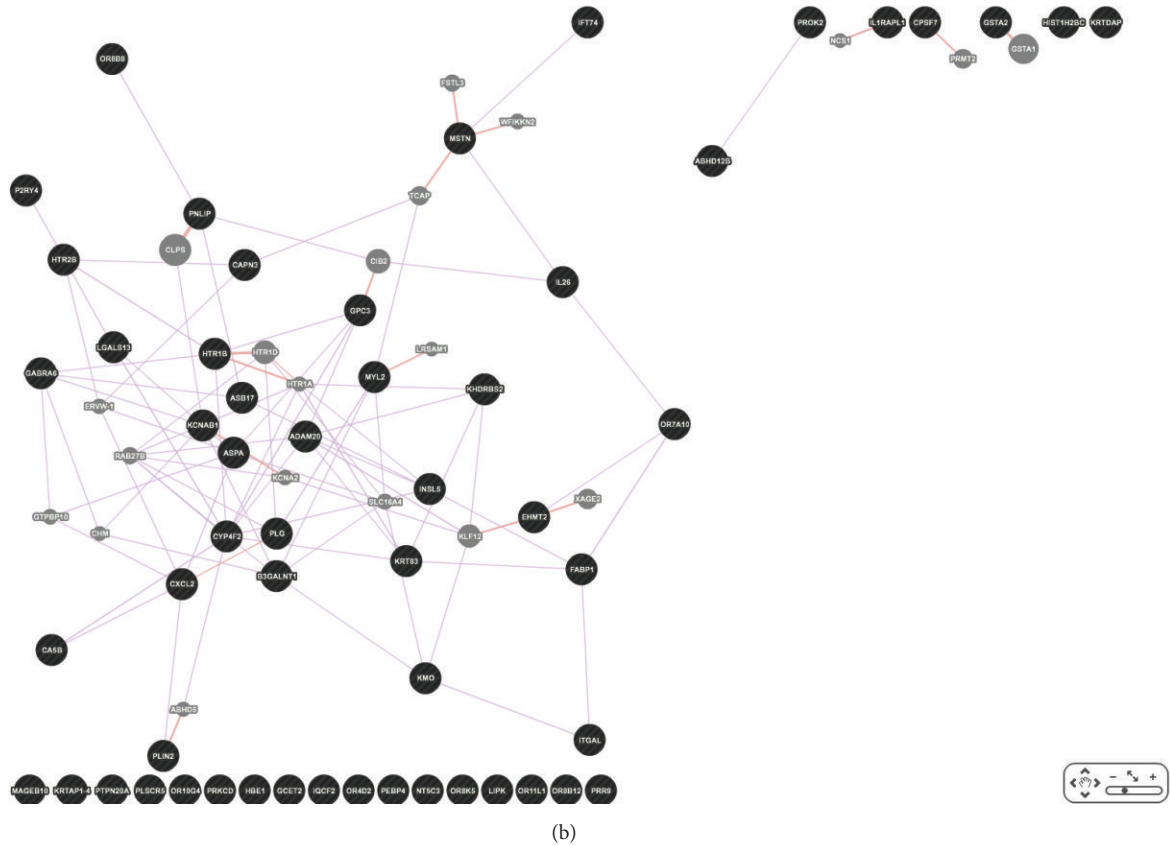


FIGURE 5: (a) Coexpression patterns from gene clustering in a sheep model data. 200 HMM-specific TDE genes are represented in heatmap. Each row contains a vector of time series expression profile in log<sub>2</sub> scale; consequently the visualization in heatmap is originally made up of major three groups, high, moderate, and low expression levels with genes that are not detected by static methods but detected by HMM, of which we selected the most statistically significant 200 genes to present this heatmap. Interestingly, some genes at low expression levels were obviously differentially expressed at log<sub>2</sub>-scaled FC ~4 up to 5 and even some genes that significantly show temporal patterns at high expression levels were also detected, yet those genes were not detected by existing static methods suggesting that HMM method reassuringly has higher sensitivity and robustness than other existing static methods in identification of differential expression regardless of expression levels. (b) Gene functional pathway and network analysis with 528 HMM specific TDE genes in a sheep model data. To explore biological functions in this gene set further, whether or not those are genuinely differential expression or random noise by chance in terms of biological insights, gene ontology (GO) and KEGG pathway analysis were performed to identify meaningful functionalities and some meaningful functions related to developmental process (intermediate mesoderm formation, regulation of cell growth involved in regulation of muscle adaptation, intermediate mesoderm formation, etc.) and gender specific terms (granulosa cell development and maternal placenta development) are detected as we anticipated to confirm the sensitivity of dynamic HMM method. The purple and pink legends represent coexpression and physical interactions across genes, respectively, and black nodes are query genes in networks.

conditions across time points. Thus, it is well known that genes work collaboratively together in a structured biological network; these biological phenomena underscore the importance taking into account the multivariate techniques when modeling temporal dynamic gene expression. Since it is not known beforehand which gene features are connected to each other, investigators sought to define informative relationships between individual gene patterns to identify many relevant classes of dynamic temporal gene expression patterns. We explored highly correlated relationships between temporal gene sets detected by bivariate dynamic methods. Pairs of trajectories were further investigated to explore the coupled coordinated relationships between different temporal patterns based on the three dependency metrics in Section 2.5. Significance levels of such relationships were estimated by

bootstrapping resampling. The methodologies to test any coupled relationship to pairs of district gene expressions are based on (1) Granger causality, (2) correlation-basis approach, and (3) Glass-s-d score as defined in [60]. In order to efficiently identify copaired temporal dynamics, using zebrafish data, we first identified statistically significant TDE genes and ran Glass-d-score based on gene permutations. Figure 8 demonstrates coupled temporal dynamics with log-scaled expression level.

3.5. Gene Functional Pathway and Network Analysis. Once temporal dynamics in gene-by-gene test and in gene-to-gene interaction were determined, the resultant temporal gene expression sets detected by ranking individual

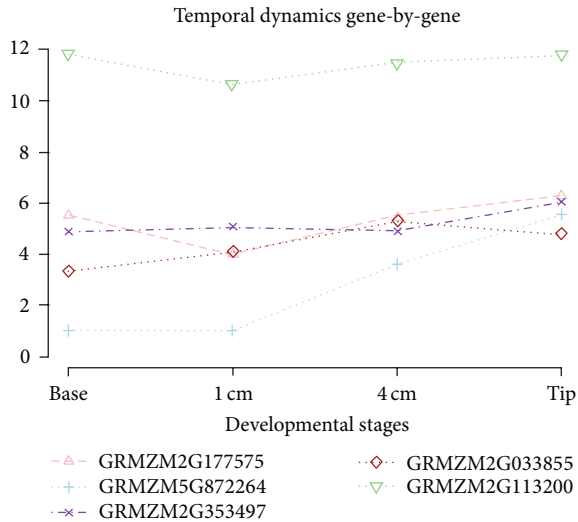


FIGURE 6: Gene expression curves for five significant TDE genes in maize leaf data. Mean expression curves are presented from two replicates during four developmental stages comparing temporal patterns each other in identifying statistically significant trajectories.

analysis and multivariate approaches, respectively, were further explored to reveal temporal relationships underlying biological processes based on gene ontology and functional network/pathway analysis. Sheep gene symbols with 21,865 genes were converted into human gene symbols with 15,343 using BioMart in R package [61]. In gene ontology analysis through Avadis NGS [62], 528 specific genes which were detected by uniquely HMM were further analyzed. Interestingly, as 63 females were sampled in the data, some of gender-specific GO terminologies among significant ontology terms were identified, that is, granulosa cell development and maternal placenta development as well as intermediate mesoderm formation, regulation of cell growth involved in cardiac muscle cell development, positive regulation of striated muscle contraction, response to stimulus involved in regulation of muscle adaptation, intermediate mesoderm formation, voluntary musculoskeletal movement, growth plate cartilage development, extracellular matrix, and so forth. The HMM-specific temporal dynamic gene sets were further investigated for coexpressed gene sets and functional network modules through ebdbNet and GeneNet in R package and GeneMANIA [63–65] as shown in Figures 5 and 6.

**3.6. Simulation Studies.** We show that dynamic methods outperform approaches that do not explicitly address the time series nature of the data in simulation studies for validation and evaluation. We evaluated the performance of dynamic methods with simulation studies in which temporal features are already known as gold standard TDE (temporally differentially expressed) gene lists. Gold standard gene lists contain entire information to mimic RNA-seq time series profile if a gene is differentially expressed (DE) or equally expressed (EE) over time as reference set to compare to the results obtained from both dynamic and static methods in terms of recall and precision measurement. To this end,

we generated simulated RNA-seq datasets with expression profiling data points representing nondifferentially expressed and differentially expressed genes in a series of time points by using different values of autocorrelation parameter ( $\phi$ ). We generated data for equally expressed genes by sampling time series process parameters ( $w$ ) of a gene in invertible Gaussian ARIMA process with  $\phi = 0$ . We generated data for differentially expressed genes across time points in the same procedures as  $\phi = 0.1, 0.25, 0.5, 0.75,$  and  $0.9$ , respectively. After time series process, regression effects and autocorrelation parameters were simulated for 1000 genes, 4 simulated datasets were generated by setting the varying number of time points and replicates in a time point,  $nT = 5$  and  $10$ ,  $nR = 3$ , and  $5$  to compute  $P$  value, FDR, and credible interval of each gene for static and dynamic methods and compared to gold standards to obtain true discovery rates in our simulated datasets.

## 4. Conclusion and Discussion

We first performed pairwise comparisons using two simple static methods, Audic-Claverie statistics and Fisher's exact test. The congruent set of both of them is highly overlapped and we reported the results of Fisher's exact test as more common method in Figure 1. The dataset came from a sheep model with two different healing systems at four different days. This dataset provides an excellent design for identification of temporally and simultaneously differentially expressed (TDE) genes as we have two conditions at each time. This type of time course is referred to as factorial time course experimental design. The authors of [53] took a union set of all these combinations of pairwise comparisons in condition and time point to identify TDE genes. These approaches might provide insights and intuitively simple static methods are alternative in small experiments in general. Evidently, the methods for time series dynamics are still in their infancy. However, those algorithms all do not consider dependency between samples in time course and they assume that all samples are independently distributed, though sequential correlation is obviously observed in data as shown in Figure 2. We noticed that basically patterns of detection of temporal changes by static and dynamic method are different, albeit they agree in some degree. That is, most of temporal genes at low and moderate expression levels are detected as significant genes in dynamic methods, whereas, due to power issues of parametric static pooling methods and simplification of pairwise methods, static methods do a good job at high expression levels. To confirm robustness and reliability of gene detection methods in time series, a comprehensive comparison and evaluation with varying parameter settings closer to RNA-seq real world is further needed. At low and moderate levels, many genes which were not detected by static methods but dynamic methods still showed log<sub>2</sub>-scaled FC  $\sim 4$  up to  $5$ . We sought to test the ability of dynamic methods whether or not those identified dynamic-unique TDE genes are genuinely differential expression or just by a random chance because expressions have been more affected by noise at low and moderate

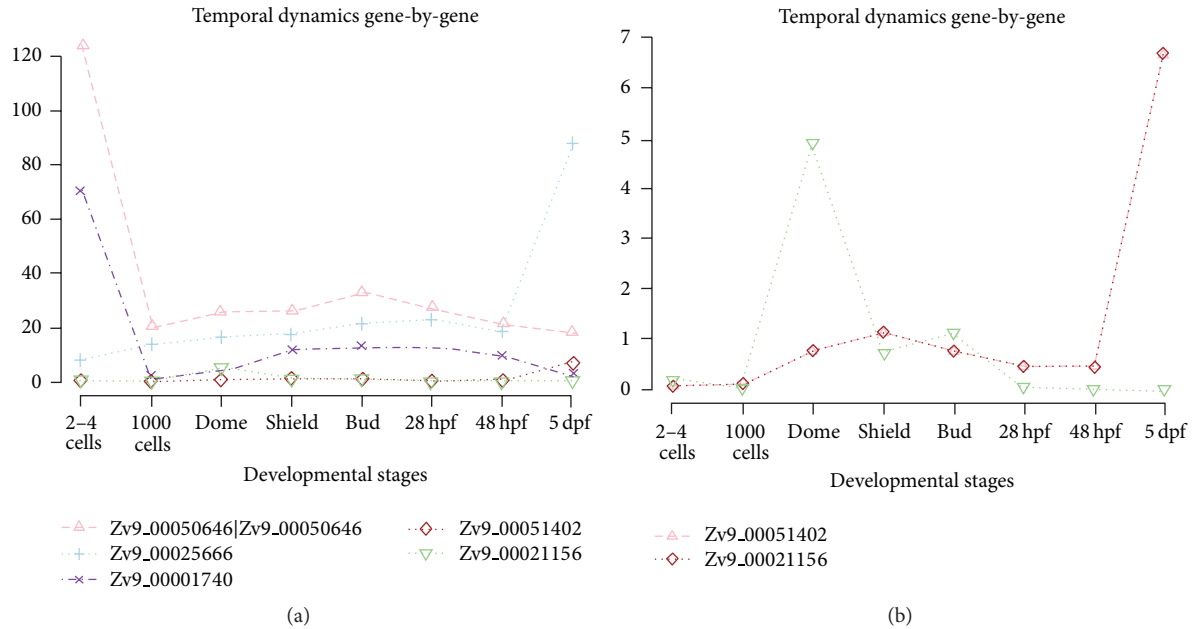


FIGURE 7: (a) Gene expression curves for five TDE genes in zebrafish data. Expression curves are presented during eight developmental stages comparing temporal patterns to each other in identifying statistical significant trajectories. (b) Two specific genes in gene-by-gene temporal dynamics with low expression levels via SETI in zebrafish data from (a). SETI enables identification of significant temporal patterns at low expression levels which are not detected by other existing static methods.

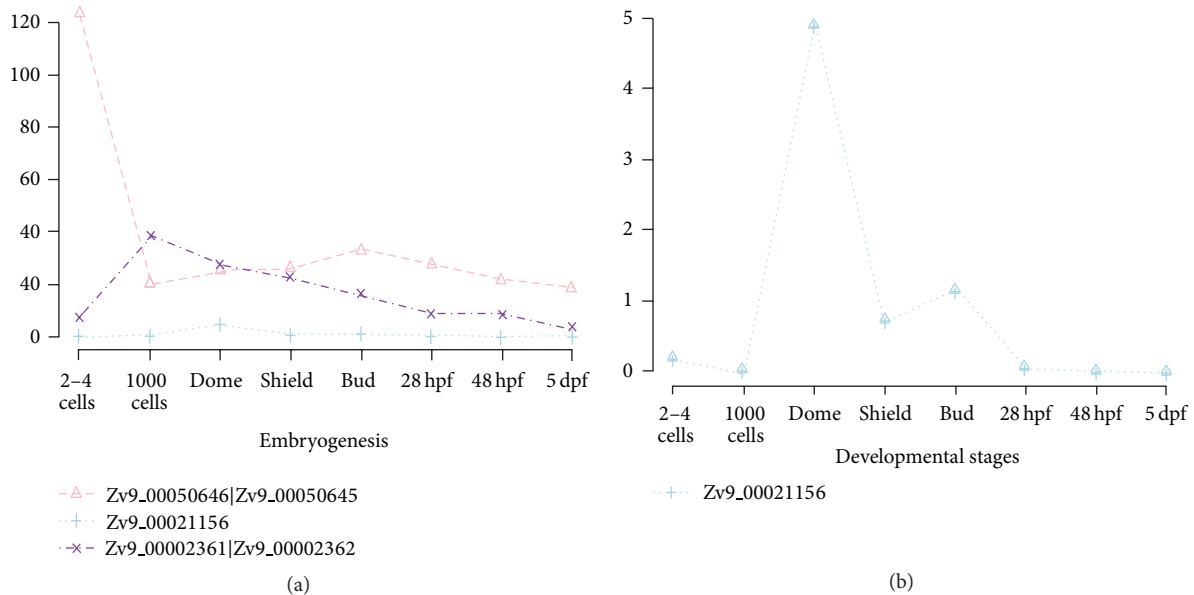


FIGURE 8: (a) A set of top cooperative TDE genes detected by Glass-d-score with FDR 0.05 as a candidate of cooperative gene pairs in coexpression in zebrafish data. Glass-d-score and corresponding FDR at cutoff 0.05 by resampling procedure under the null hypothesis that there are no TDE patterns across samples and those are shuffled with 1,000 repetitions. (b) One specific gene in coupled temporal dynamics with low expression level via SETI and Glass-d-score from (a). Glass-d-score is robust in identifying coexpressed genes over time at low expression levels.

levels in microarray, even though RNA-seq quality when compared to microarray has been much improved for now. In Figure 4, to assess temporally differentially expressed genes, we incorporated SETI with HMM algorithm in a sheep model within each condition to see a variety of time-varying

trajectories since HMM provides only patterns of hidden latent variables (DE/EE) at developmental stages. Left panel shows three candidate genes at low and moderate levels and right panel shows another three candidate genes at high expression. To examine biological meanings in TDE

genes detected by only HMM, we further performed gene clustering coexpression patterns to see if those gene sets have possibility of false negatives in altered expression of cooperative genes and gene functional pathway analyses. Notably, we confirmed meaningful biological functionalities and temporal patterns in dynamic specific TDE genes in downstream analyses, gene clustering, gene ontology, and pathway/network analysis. Interestingly, as 63 females were sampled in the data, some of gender-specific GO terminologies among significant ontology terms were identified, that is, granulosa cell development and maternal placenta development, intermediate mesoderm formation, regulation of cell growth involved in cardiac muscle cell development, positive regulation of striated muscle contraction, response to stimulus involved in regulation of muscle adaptation, intermediate mesoderm formation, voluntary musculoskeletal movement, growth plate cartilage development, extracellular matrix, and so forth. Consistently, HMM, SETI, and AR(1) model that account for time dependency Markovian property in the models identified more of statistically significant TDE genes than static methods regardless of expression levels. In summary, the approaches we described use a developed unified dynamic test framework that includes SETI with statistical significance testing, ranking temporal genes by AR(1) modeling and posterior probability of autocorrelation parameter, and HMM to classify temporal dynamic patterns. These methods seem to be robust regardless of the magnitude of expression (see Figures 7(b) and 8(b), and Supplementary Tables) and more sensitive than static methods as shown in Supplementary TDE Tables; moreover, TDE genes detected by dynamic specific methods were confirmed as temporal dynamics in clustering patterns and biologically significant modules in network analysis implying that the gene sets were not identified as false negative genes in static methods that samples over time are assumed independently. We anticipate that temporal RNA-seq experiments will be widely performed in the near future due to reduced sequencing cost and the rich information carried by these experiments. In this paper, we consider several statistical approaches that can explicitly model the time series nature in the data. We discussed the limitations of simple static pairwise comparison methods for time series data analysis; dynamic statistical framework for RNA-seq read count with statistical evolutionary trajectory index measure; autoregressive time-lagged AR(1) model; hidden Markov model; pairwise and multiple comparisons among trajectories to investigate coupled bivariate dependency between distinct SETIs; and pathway/network analysis in transcriptome data based on detected temporally differentially expressed genes. Thus, this study covers critical issues that have not been systematically addressed in temporal RNA-seq data and we hope this will motivate more rigorous developments of novel methods to model and analyze RNA-seq data. Of particular interest will be the extension of these methods to combined time series datasets from RNA-seq, proteomics, and metabolomics for *in silico* cell/organism predictive modeling [6]. In addition, it will facilitate cross-species comparative analyses of temporal gene expression to investigate developmental processes and disease progression such as aging and virus-mediated

immune disease dynamics. Deep sequencing of mRNAs has been a popular and effective approach for quantification of alternative splicing events, and it is well known that more than 90 percent of human genes have multiple isoforms to produce different protein structures. Thus, an important future direction is also to extend the statistical framework of our dynamic methods to incorporate the characterization of isoform diversity in time course in detecting differential expression.

## Acknowledgments

This work was supported in part by NIH GM094780 (J. P. Noonan), GM59507 (H. Zhao) and NSF DMS 1106738 (H. Zhao).

## References

- [1] S. Marguerat and J. Bähler, “RNA-seq: from technology to biology,” *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.
- [2] B. A. Friedman and T. Maniatis, “ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data,” *Genome Biology*, vol. 12, p. R69, 2011.
- [3] D. Ghosh and Z. S. Qin, “Statistical issues in the analysis of ChIP-seq and RNA-seq data,” *Genes*, vol. 1, no. 2, pp. 317–334, 2010.
- [4] L. M. McIntyre, K. K. Lopiano, A. M. Morse et al., “RNA-seq: technical variability and sampling,” *BMC Genomics*, vol. 12, p. 293, 2011.
- [5] Z. Fang and X. Cui, “Design and validation issues in RNA-seq experiments,” *Briefings in Bioinformatics*, vol. 12, no. 3, Article ID bbr004, pp. 280–287, 2011.
- [6] J. R. Karr, J. C. Sanghvi, D. N. Macklin et al., “A whole cell computational model predicts phenotype from genotype,” *Cell*, vol. 150, no. 2, pp. 389–401, 2012.
- [7] S. M. Wang, “Understanding SAGE data,” *Trends in Genetics*, vol. 23, no. 1, pp. 42–50, 2007.
- [8] A. Schliep, A. Schönhuth, and C. Steinhoff, “Using hidden Markov models to analyze gene expression time course data,” *Bioinformatics*, vol. 19, no. 1, pp. i255–i263, 2003.
- [9] R. Gottardo, A. E. Raftery, K. Yee Yeung, and R. E. Bumgarner, “Bayesian robust inference for differential gene expression in microarrays with multiple samples,” *Biometrics*, vol. 62, no. 1, pp. 10–18, 2006.
- [10] X. Han, W. K. Sung, and L. Feng, “Pem: a general statistical approach for identifying differentially expressed genes in time-course cDNA microarray experiment without replicate,” in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, pp. 123–132, 2006.
- [11] Y. C. Tai and T. P. Speed, “A multivariate empirical Bayes statistic for replicated microarray time course data,” *Annals of Statistics*, vol. 34, no. 5, pp. 2387–2412, 2006.
- [12] Y. C. Tai and T. P. Speed, “On gene ranking using replicated microarray time course data,” *Biometrics*, vol. 65, no. 1, pp. 40–51, 2009.
- [13] P. Ma, W. Zhong, and J. S. Liu, “Identifying differentially expressed genes in time course microarray data,” *Statistics in Biosciences*, vol. 1, no. 2, pp. 144–159, 2009.

- [14] Y. Goltsev and D. Papatsenko, "Time warping of evolutionary distant temporal gene expression data based on noise suppression," *BMC Bioinformatics*, vol. 10, p. 353, 2009.
- [15] M. J. Aryee, J. A. Gutiérrez-Pabello, I. Kramnik, T. Maiti, and J. Quackenbush, "An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation)," *BMC Bioinformatics*, vol. 10, p. 409, 2009.
- [16] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12837–12842, 2005.
- [17] T. Park, S. G. Yi, S. Lee et al., "Statistical tests for identifying differentially expressed genes in time-course microarray experiments," *Bioinformatics*, vol. 19, no. 6, pp. 694–703, 2003.
- [18] W. Xu, J. Seok, M. N. Mindrinos et al., "Human transcriptome array for high-throughput clinical studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 9, pp. 3707–3712, 2011.
- [19] Y. Liang, B. Tayo, X. Cai, and A. Kelemen, "Differential and trajectory methods for time course gene expression data," *Bioinformatics*, vol. 21, no. 13, pp. 3009–3016, 2005.
- [20] A. A. Kalaitzis and N. D. Lawrence, "A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression," *BMC Bioinformatics*, vol. 12, p. 180, 2011.
- [21] J. J. Song, H. J. Lee, J. S. Morris, and S. Kang, "Clustering of time-course gene expression data using functional data analysis," *Computational Biology and Chemistry*, vol. 31, no. 4, pp. 265–274, 2007.
- [22] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.
- [23] I. Sohn, K. Owzar, S. L. George, S. Kim, and S. H. Jung, "A permutation-based multiple testing method for time-course microarray experiments," *BMC Bioinformatics*, vol. 10, p. 336, 2009.
- [24] S. C. Billups, M. C. Neville, M. Rudolph, W. Porter, and P. Schedin, "Identifying significant temporal variation in time course microarray data without replicates," *BMC Bioinformatics*, vol. 10, p. 96, 2009.
- [25] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, no. 4, pp. 507–508, 2006.
- [26] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nature Genetics*, vol. 32, no. 5, pp. 490–495, 2002.
- [27] J. B. Fan, M. S. Chee, and K. L. Gunderson, "Highly parallel genomic assays," *Nature Reviews Genetics*, vol. 7, no. 8, pp. 632–644, 2006.
- [28] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, p. 191, 2006.
- [29] Y. Yuan, Y. P. Chen, S. Ni et al., "Development and application of a modified dynamic time warping algorithm (DTW-S) to analyses of primate brain expression time series," *BMC Bioinformatics*, vol. 12, p. 347, 2011.
- [30] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola, "Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 18, pp. 10146–10151, 2003.
- [31] Y. Yuan, C. T. Li, and R. Wilson, "Partial mixture model for tight clustering of gene expression time-course," *BMC Bioinformatics*, vol. 9, p. 287, 2008.
- [32] G. C. Tseng and W. H. Wong, "Tight clustering: a resampling-based approach for identifying stable and tight patterns in data," *Biometrics*, vol. 61, no. 1, pp. 10–16, 2005.
- [33] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.
- [34] T. T. Nguyen, R. R. Almon, D. C. DuBois, W. J. Jusko, and I. P. Androulakis, "Importance of replication in analyzing time-series gene expression data: corticosteroid dynamics and circadian patterns in rat liver," *BMC Bioinformatics*, vol. 11, p. 297, 2010.
- [35] M. J. Nueda, J. Carbonell, I. Medina, J. Dopazo, and A. Conesa, "Serial expression analysis: a web tool for the analysis of serial gene expression data," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq488, pp. W239–W245, 2010.
- [36] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
- [37] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.
- [38] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.
- [39] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, 2010.
- [40] 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 Research*, vol. 18, no. 9, pp. 1509–1517, 2008.
- [41] S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferre, and A. Conesa, "Differential expression in RNA-Seq: a matter of depth," *Genome Research*, vol. 21, pp. 2213–2223, 2011.
- [42] T. J. Hardcastle and K. A. Kelly, "BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, p. 422, 2010.
- [43] P. L. Auer and R. W. Doerge, "Statistical design and analysis of RNA sequencing data," *Genetics*, vol. 185, no. 2, pp. 405–416, 2010.
- [44] P. L. Auer, S. Srivastava, and R. W. Doerge, "Differential expression—the next generation and beyond," *Briefings in Functional Genomics*, 2011.
- [45] J. Lee, Y. Ji, S. Liang, G. Cai, and P. Muller, "On differential gene expression using RNA-Seq data," *Cancers Information*, vol. 10, pp. 205–215, 2011.
- [46] A. Oshlack, M. D. Robinson, and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biology*, vol. 11, no. 12, p. 220, 2010.
- [47] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

- [48] M. Yuan and C. Kendzierski, "Hidden Markov models for microarray time course data in multiple biological conditions," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1323–1332, 2006.
- [49] M. Jäger, C. E. Ott, J. Grünhagen et al., "Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing," *BMC Genomics*, vol. 12, p. 158, 2011.
- [50] P. Li, L. Ponnala, N. Gandotra et al., "The developmental dynamics of the maize leaf transcriptome," *Nature Genetics*, vol. 42, no. 12, pp. 1060–1067, 2010.
- [51] A. Pauli, E. Valen, M. F. Lin et al., "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis," *Genome Research*, vol. 22, pp. 577–591, 2012.
- [52] J. Lienau, K. Schmidt-Bleek, A. Peters et al., "Insight into the molecular pathophysiology of delayed bone healing in a sheep model," *Tissue Engineering A*, vol. 16, no. 1, pp. 191–199, 2010.
- [53] H. Schell, M. S. Thompson, H. J. Bail et al., "Mechanical induction of critically delayed bone healing in sheep: radiological and biomechanical results," *Journal of Biomechanics*, vol. 41, no. 14, pp. 3066–3072, 2008.
- [54] J. L. Hay and A. N. Pettitt, "Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease," *Biostatistics*, vol. 2, no. 4, pp. 433–444, 2011.
- [55] M. L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 18, pp. 9834–9839, 2000.
- [56] C. Spyrou, R. Stark, A. G. Lynch, and S. Tavaré, "BayesPeak: Bayesian analysis of chip-seq data," *BMC Bioinformatics*, vol. 10, p. 299, 2009.
- [57] M. A. Newton, C. M. Kendzierski et al., *Parametric Empirical Bayes Methods for Microarrays in the Analysis of Gene Expression Data: Methods and Software*, Springer, New York, NY, USA, 2003.
- [58] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*, E. Beckenbach, Ed., McGraw-Hill, New York, NY, USA, 1956.
- [59] S. Audic and J. M. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.
- [60] G. Glass, "Integrating findings: the meta-analysis of research," *Review of Research in Education*, vol. 5, pp. 351–379, 1977.
- [61] "bioMart in bioconductor," <http://www.bioconductor.org/>.
- [62] Strand Life Sciences Pvt. Ltd., *Avadis NGS Version: Version 1.3.0*, Strand Scientific Intelligence, San Francisco, NC, USA, 2012.
- [63] D. Warde-Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq537, pp. W214–W220, 2010.
- [64] "GeneNet in bioconductor," <http://www.bioconductor.org/>.
- [65] "ebdbNet in bioconductor," <http://www.bioconductor.org/>.