Modeling Expression Plasticity of Genes that Differentiate Drug-sensitive from Drug-resistant Cells to Chemotherapeutic Treatment

Ningtao Wang[1,2], Yaqun Wang[1,2], Hao Han[1,2], Kathryn J. Huber[3], Jin-Ming Yang[3], Runze Li[1,2] and Rongling Wu[2,1,*]

[1]*Department of Statistics, Pennsylvania State University, State College, PA 16802, USA; [2]Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA; [3]Department of Pharmacology, The Pennsylvania State University, Hershey, PA 17033, USA*

Abstract: By measuring gene expression at an unprecedented resolution and throughput, RNA-seq has played a pivotal role in studying biological functions. Its typical application in clinical medicine is to identify the discrepancies of gene expression between two different types of cancer cells, sensitive and resistant to chemotherapeutic treatment, in a hope to predict drug response. Here we modified and used a mechanistic model to identify distinct patterns of gene expression in response of different types of breast cancer cell lines to chemotherapeutic treatment. This model was founded on a mixture likelihood of Poisson-distributed transcript read data, with each mixture component specified by the Skellam function. By estimating and comparing the amount of gene expression in each environment, the model can test how genes alter their expression in response to environment and how different genes interact with each other in the responsive process. Using the modified model, we identified the alternations of gene expression between two cell lines of breast cancer, resistant and sensitive to tamoxifen, which allows us to interpret the expression mechanism of how genes respond to metabolic differences between the two cell types. The model can have a general implication for studying the plastic pattern of gene expression across different environments measured by RNA-seq.

Received on: June 24, 2014 - Revised on: August 4, 2014 - Accepted on: August 24, 2014

Keywords: RNA-seq, Phenotypic plasticity, Gene-environment interaction, Clustering.

INTRODUCTION

An organism's ability to adapt to changes in the environment, called phenotypic plasticity, is essential for the survival of the organism [1, 2]. Occurring at all levels of biological organization from cells to organisms, plastic response to various internal and external environmental signals represents an intrinsic attribute of organisms that facilitates evolution [3, 4]. In recent years, the concept of phenotypic plasticity has been increasingly used to study the causes of complex human diseases [5-8]. It has been recognized that the formation of phenotypic plasticity is mediated through altered patterns of gene and protein expression cued by the environment [9-12], but the mechanistic details of cell and molecular biology behind phenotypic plasticity are not fully understood.

Current next-generation sequencing techniques (RNA-seq) provide a powerful tool to array the expression of whole-genome transcript genes [13, 14]. By linking these genes to environmental variation, one can identify the pattern of how genes alter their expression to cause phenotypic plasticity when the environment changes [15]. In practice, numerous genes have been identified to be up- or down-regulated in response to specific environmental signals [16, 17]. Traditional approaches for identifying environment-induced gene differentiation are based on a simple comparative analysis of expression for individual genes between different environments. Other approaches cluster genes into different groups based on their biological function under one specific condition or treatment [18]. No approaches are currently available to catalogue genes in terms of their pattern of expression in response to environmental stimuli, thereby limiting our inference about the mechanisms governing differential expression of genes across environments.

More recently, Wang *et al.* [19] developed a bi-variate Poisson model to cluster genes expressed in two different environments. Jiang *et al.* [20] derived an algorithmic model for clustering genes based on their environment-induced differentiation patterns. Both models incorporate distinct patterns of gene expression by RNA-seq into a mixture likelihood framework. Each mixture component of the likelihood corresponds to a particular expression pattern that distinguishes this component from the other. By integrating intrinsic environment-dependent plasticity, the discoveries of expression patterns from Jiang *et al.*'s model are biologically more interpretable than those from traditional clustering approaches using a single environment.

The main idea of Jiang *et al.*'s model is to cluster the differences of gene expression between two treatments, making it possible to characterize the patterns of gene differentiation directly from environmental influences on transcription. How-

[*]Address correspondence to this author at the Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA; Tel: +001 717 531 2037; Fax: +001 717 531 0480; E-mail: rwu@phs.psu.edu.

1875-5488/14 $58.00+.00

ever, the difference of two Poisson variables follows a complex form of distribution which makes parameter estimation highly challenging [21, 22]. We modified Jiang *et al.*'s model by integrating an integrative generalized EM and Newton-Raphson algorithm to estimate the parameters of gene expression that describe transcriptional plasticity to changing environment. We used the new modified model to analyze RNA-seq data that describe global changes of gene expression between two types of breast cancer cell lines, resistant and sensitive to the intervention of tamoxifen, leading to the identification of differential expression patterns in response to metabolic differences between the two cell types. Computer simulation was performed to examine the statistical properties of the new model and validate its usefulness and utilization.

MODEL

Mixture

Suppose we have measured the expression reads of n genes in two different treatments, such as different tissues, different cell types, or different temperatures. Let m_{1i} and m_{2i} denote the expression reads of gene i measured in the two treatments, respectively. The difference of expression of this gene between the treatments is calculated as $m_i = m_{1i} - m_{2i}$. This difference is used as a measure of phenotypic plasticity [1]. Because of their functional similarities and differences in plastic response, these genes can be clustered into different groups (assuming J groups). Thus, for any gene i, it should arise from one (and only one) of the J groups. The likelihood of the expression data of n genes is written as

$$L(\Theta \mid m) = \prod_{i=1}^{n}\left[\pi_1 p_1(m_i) + \ldots + \pi_J p_J(m_i)\right] \qquad \textbf{(1)}$$

where Θ is a set of unknown parameters; $(\pi_1, \ldots, \pi_J)$ is a set of proportions that each corresponds to a gene group; and $p_j(m_i)$ is the discrete probability distribution of differential expression for group j. The expression reads of genes in each treatment are thought to obey a Poisson distribution [19], thus the distribution of the read differences between the two cell types is modeled by the Skellam function [22, 23], expressed as

$$p_j(m_i) = e^{-(\lambda_{1j}+\lambda_{2j})}\left(\frac{\lambda_{1j}}{\lambda_{2j}}\right)^{\frac{m_i}{2}} I_{|m_i|}\left(2\sqrt{\lambda_{1j}\lambda_{2j}}\right), \forall j = 1, \ldots, J \; \textbf{(2)}$$

where λ_{1j} and λ_{2j} are the expected numbers of reads for all genes that belong to group j in two treatments, respectively; and $I_{|m_i|}\left(2\sqrt{\lambda_{1j}\lambda_{2j}}\right)$ is the modified Bessel function of the first kind, which is the solution of Bessel's differential equation [21]. The modified Bessel function of the first kind is expressed as

$$I_{|m_i|}\left(2\sqrt{\lambda_{1j}\lambda_{2j}}\right) = \sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,\Gamma\left(k+|m_i|+1\right)}\left(\sqrt{\lambda_{1j}\lambda_{2j}}\right)^{2k+|m_i|} \quad \forall j = 1, \ldots, J,$$

where $\Gamma\left(k+|m_i|+1\right)$ is the gamma function, a generalization of the factorial function to non-integer values. All unknown parameters are arrayed in $\Theta = (\{\pi_j, \lambda_{1j}, \lambda_{2j}\}_{j=1}^{J})$.

Parameter Estimation

To obtain the maximum likelihood estimates (MLEs) of the parameter vector, Θ, that defines the mixture model (1), we implement an integrative procedure that combines the generalized EM and Newton-Raphson algorithms for parameter estimation. From the mixture (1), we define the E step by calculating the posterior probability at which a gene i belongs to group j given the observations and parameter estimates. Using the results from the E step, we obtain the M step in which the proportion of group j is calculated using a closed form estimator and the expected values of gene group j in the two treatments, λ_{1j} and λ_{2j}, are calculated using the Newton-Raphson algorithm. A detailed procedure for parameter estimation is given in the Appendix.

For a practical data set, we do not know the optimal number of gene clusters. This can be determined by a model selection criterion, such as commonly used Akaike information criterion (AIC) or Bayesian information criterion (BIC). The optimal number of clusters corresponds to the minimum AIC or BIC value.

Hypothesis Tests

After the parameters are estimated, we will perform two biologically meaningful tests as follows:

Test 1: For a given cluster group, genes are differently expressed between the two treatments. This can be done by testing:

$$H_0: \lambda_{1j} = \lambda_{2j} \text{ vs. } H_1: \lambda_{1j} \neq \lambda_{2j} \;\forall j = 1, \ldots, J \qquad \textbf{(3)}$$

If the H_0 is rejected, this group of genes displays different amounts of expressions between the two treatments, indicating that they may contribute to phenotypic plasticity and can be viewed as a predictor of this phenomenon.

Test 2: For a pair of gene groups j and k, they interact with each other to determine phenotypic plasticity. This can be done by testing:

$$H_0: \lambda_{1j}\text{–}\lambda_{1k} = \lambda_{2j}\text{–}\lambda_{2k} \text{ vs.}$$

$$H_1: \lambda_{1j}\text{–}\lambda_{1k} \neq \lambda_{2j}\text{–}\lambda_{2k} \;\forall j < k = 1, \ldots, J. \qquad \textbf{(4)}$$

A rejection of the H_0 means that these two groups of genes have significant interaction effects with treatment.

Test 3: The extent of phenotypic plasticity is often associated with the magnitude of change of environmental signals. Thus, it is interesting to test whether the change of gene expression for a particular group is consistent with the extent of change of the environment. This can be done by formulating the hypotheses:

$$H_0: \lambda_{1j}/\lambda_{2j} = c \text{ vs. } H_1: \lambda_{1j}/\lambda_{2j} \neq c \;\forall j = 1, \ldots, J \qquad \textbf{(5)}$$

where c is the ratio of the environmental signals between the two treatments.

The log-likelihood ratio (LR) test statistics for each of the three hypotheses (3) – (5) are calculated. The LR values are approximated by a chi-square distribution with the degree of freedom equaling the difference of the number of parameters to be estimated under the H_0 and H_1.

Consideration of Overdispersion

Many clustering models assume that RNA-seq reads follows a Poisson distribution, although several recent studies do not support this assumption because of overdispersion [24-26]. In statistics, the overdispersion can be captured by a negative binomial (NB) distribution, in which reads follow a Poisson distribution, but with the mean approximated by a Gamma distribution. However, our model is based on differences of reads between two treatments, which have become a finite Poisson (fPMD) mixture distribution, rather than a simple Poisson distribution as assumed for reads in individual treatments by the previous models [19]. In the fPMD, the mean is modeled as a finite discrete distribution, which can well take into account the overdispersion of across-treatment differences of reads.

To support our argument, we simulated two sets of read data, each with 400 genes from NB(5, 0.5) and NB(10, 0.5), respectively, allowing overdispersion. We then calculated differences between the same genes from these two sets. As shown in (Fig. **1**), the differences can well be fitted by the Skellam mixture model. From this example, the Skellam mixture distribution can be safely used to model the distribution of differences of reads between two treatments, even if overdispersion occurs for these reads.

RESULTS

The newly developed model was used to analyze a real data set on the phenotypic plasticity of gene expression. As a commonly used adjuvant hormonal therapy for patients with breast cancer, tamoxifen blocks the effects of estrogen in breast cancer cells by mediating the estrogen receptor (ER) to prevent ER-mediated transcription. Although tamoxifen has successfully treated some ER-negative breast tumors [27], its efficacy has often been limited by drug resistance [28]. To reveal the global mechanisms of gene expression and signaling pathway alterations for tamoxifen resistance, Huber-Keener *et al.* [29] compared the transcriptomes of breast cancer cells that are tamoxifen-sensitive and tamoxifen-resistant by collecting a total of 23,561 mRNA genes using RNA-Seq.

As the demonstration of how the model can be used in practice, we randomly chose 500 from this set of genes. By using the differences of gene expression between tamoxifen-sensitive and tamoxifen-resistant cell types, we used the Skellam model to analyze the data, aimed to detect genes that are associated to the development of tamoxifen resistance. Based on the BIC values under different numbers of clusters, 500 randomly chosen genes are categorized into four distinct clusters (Fig. **2**). The proportions of the four clusters were estimated and all genes were found to be up-regulated from tamoxifen-sensitive to tamoxifen-resistant cell types, but the number of genes within a cluster decreases with the increasing extent of differentiation of gene expression between the two cell types (Table **1**). This suggests that a fewer number of genes display pronounced differentiation over the two treatments. It is interesting to note that treatment-induced differentiation of gene expression can be more precisely estimated than the amount of gene expression in each treatment, as shown by the standard errors of the estimates (Table **1**).

(Fig. **3A**) plots the patterns of how genes are differently expressed from tamoxifen-sensitive to tamoxifen-resistant cells. About a half are only slightly up-regulated (cluster 1), whereas about 15% of genes increase their expression dramatically in the resistant cells (cluster 4). The other genes (clusters 2 and 3) are up-regulated moderately from tamoxifen-sensitive to tamoxifen- resistant cell types. Hypothesis test (3) was used to examine whether each cluster of genes is expressed significantly differently between the two cell types, with the result suggesting that all clusters are significant ($p < 2.01{\times}10^{-5}$).

Based on the estimated λ_{1j} and λ_{2j} values, we drew the plots of expression against the cell type for each cluster (Fig. **3B**). Cluster 4 is not only expressed much more strongly in both cell types than the other three clusters which display a similar amount of expression, but also is the most sensitive to metabolic changes between the two cell types among all the clusters. Using hypothesis test (4), we investigated how each pair of clusters interact with cell types. Significant interactions were detected for each cluster pair; for example, cluster 1 vs. 2 at $P = 2.81 \times 10^{-6}$, cluster 2 vs. 3 at $P = 2.85 \times 10^{-6}$, and cluster 3 vs. 4 at $P = 1.28 \times 10^{-6}$.

We also tested whether the ratio of the amounts of gene expression between the two cell line types is identical among clusters. The result from this test indicates that none of cluster pairs has the same ratio of treatment-dependent gene expression. This suggests that all these clusters function differently in response to the metabolic environment of the cell types.

We compared our Skellam mixture model with a more commonly used hierarchical clustering approach with Ward's criterion [30]. Ward's criterion aims to minimize the total within-cluster variance while maximizing the between-cluster variance. By analyzing the same data, the new model

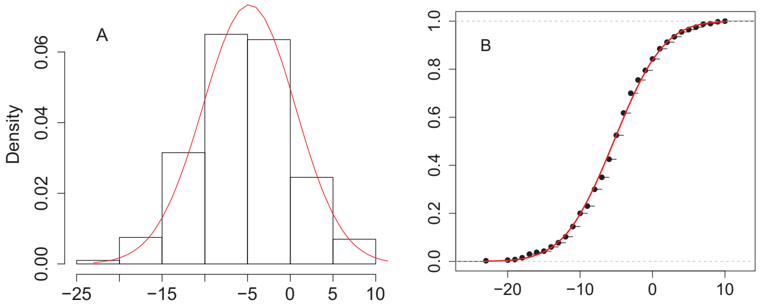

Fig. (1). Differences of reads between two treatments (black) fitted by the Skellam mixture distribution (red). (**A**) Density function expressed by a histogram and curve. (**B**) Distribution function expressed by observed points and curve.

is found to provide a similar result with that by the hierarchical model (Fig. **4**), suggesting the statistical reasonability of the new model. However, our model allows various quantitative inferences of gene differentiation as formulated in hypothesis tests (3) – (5), thereby with results from our model being biologically more interpretable, informative and implementable to practical settings than those from traditional approaches.

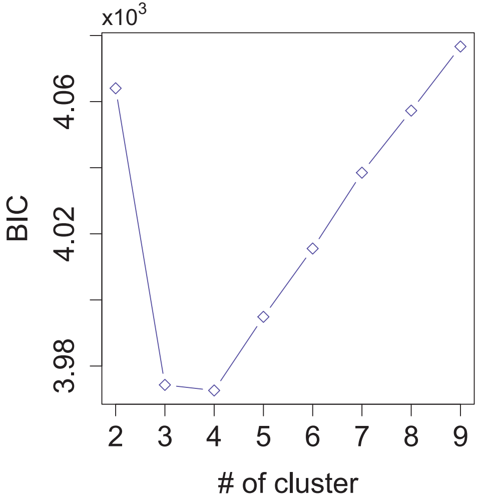

Fig. (2). BIC values calculated under an increasing number of gene clusters detected by the model. The optimal number of clusters corresponds to the minimum BIC value.

COMPUTER SIMULATION

We performed simulation studies to examine the statistical behavior of the new Skellam model by investigating the precision of parameter estimation. We simulated read data of 500 transcript genes with four distinct clusters by mimicking the tamoxifen example as described above. The model was used to analyze the simulated data and find the number of clusters. The BIC values indicate that the model correctly finds four clusters. Among 1,000 simulation replicates, over 95% can provide a correct estimate of cluster numbers.

(Table **2**) tabulates the estimates of the mean reads in two different treatments, λ_{1j} and λ_{2j}, for four clusters. It can be seen that the model provides reasonable estimates of mean read counts for each cluster and obtains better estimates of the differences of mean read counts between two treatments than mean counts in individual treatments. Using the estimated values, we drew the plots of each cluster over two treatments, in a comparison with those obtained from true values (Fig. **5**). The broad consistency between the estimated and true plots suggests that our model can provides reasonable estimates of the patterns of gene differentiation in response to environmental change. As expected, the estimates of parameters for a larger cluster are better than those for a smaller cluster. By changing the values of mean reads and their environment-dependent differences, additional simulation was carried out to investigate the influence of different parameter values on the estimation precision. In general, reasonable estimates can be obtained in all these cases, except for a small cluster with 50 genes or less.

We also carried out simulation studies by changing the sample size, the amount of gene expression in each treatment, and treatment-dependent difference of gene expression. Results from these simulation studies allow the practitioners to determine an optimal sample size under various situations of gene expression differentiation.

DISCUSSION

We have developed a new model for clustering gene expression profiles, measured by RNA-seq, based on their differentiation in response to different agents. The model is based on a mixture likelihood in which each component is specified by a particular pattern of gene expression related to a certain biological function [31]. The model fully considers the statistical feature of transcript read data by next-generation sequencing [19]; meanwhile, it displays several biological and statistical merits.

By jointly capitalizing on expression data from two treatments, our model provides more power for gene identification than conventional clustering approaches based on individual treatments [18]. The new model identifies different patterns of gene differentiation according to their plastic response to environmental change, therefore facilitating an understanding of mechanistic basis for the association between gene expression and phenotypic plasticity, a phenomenon that pervades the biological kingdom [1, 2, 32]. Given its increasing implication for studying the etiology of human diseases [5-8], there is a daunting need on the understanding of the genetic architecture of this phenomenon.

Although many previous models assume the Poisson distribution of RNA-seq data [19], some authors recently found that this assumption may not always work due to the overdispersion of data [24-26]. However, our model clusters genes into different groups by using the differences of their expression between environments. The distribution of the

Table 1. The MLEs of the mean reads for each cluster in individual treatments and their differences between the two treatments, tamoxifen-sensitive and tamoxifen-resistant cell types. The standard errors of the MLEs are also given.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Prop.	0.42±0.028	0.24±0.028	0.17±0.024	0.16±0.020
λ_{1j}	2.18±0.342	2.56±2.446	7.16±8.105	230±41.42
λ_{2j}	3.77±0.473	12.1±2.673	29.6±7.826	285±41.19
λ_{1j}– λ_{2j}	1.59±0.250	9.58±0.735	22.4+1.316	54.8±3.385

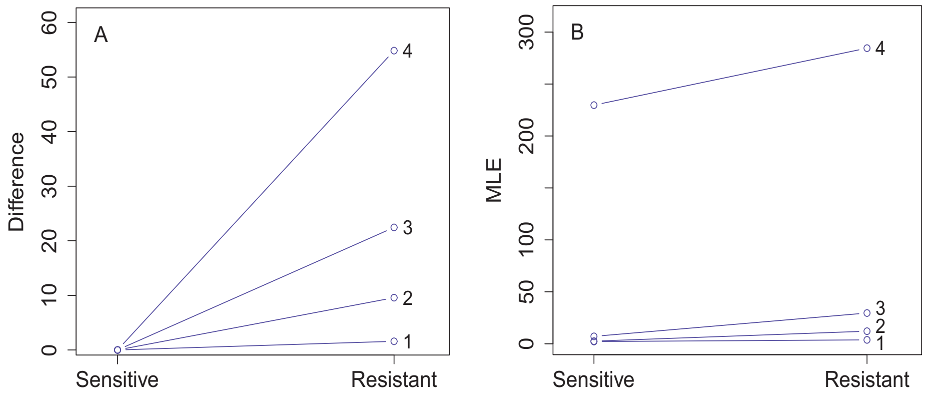

Fig. (3). Four patterns of gene expression in response to metabolic changes of tamoxifen-sensitive and tamoxifen-resistant breast cancer cell types detected by the Skellam model. (**A**) Differences of gene expression for each cluster between the two cell types are shown. (**B**) Actual expression values of four clusters are plotted over the cell types.

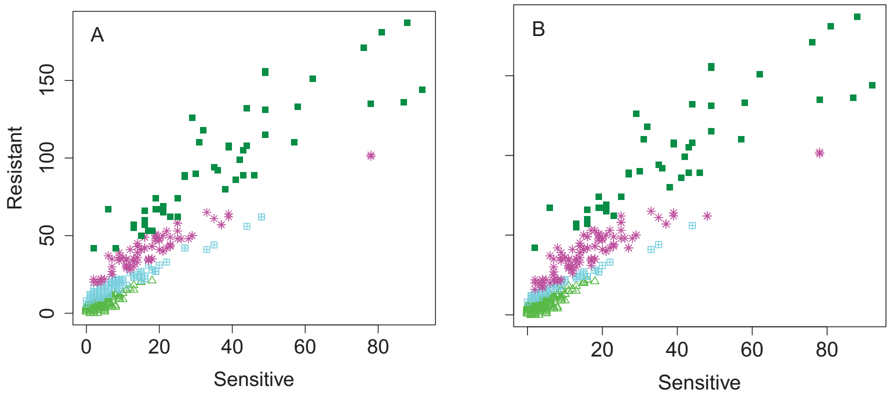

Fig. (4). Comparison of the Skellam mixture model (**A**) and hierarchical clustering model with Ward's criterion (**B**) by analyzing the tamoxifen-resistant and sensitive data of gene expression. Dots in different shapes denote four clusters detected.

Table 2. The MLEs of the mean reads for each cluster in individual treatments and their differences between the two treatments. The standard deviations (SD) of the MLEs are also given.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	True	MLE±SD	True	MLE±SD	True	MLE±SD	True	MLE±SD
Prop.	0.42	0.44±0.028	0.24	0.25±0.030	0.17	0.19±0.017	0.16	0.12±0.014
λ_{1j}	2.18	1.82±0.391	2.56	3.98±2.352	7.16	11.4±5.193	230	219±54.46
λ_{2j}	3.77	3.24±0.527	12.1	13.8±2.301	29.6	34.9±4.898	285	277±52.57
$\lambda_{1j}-\lambda_{2j}$	1.59	1.43±0.249	9.58	9.84±0.762	22.4	23.4+1.149	54.8	57.8±4.240

differences of two Poisson variables, which is a finite Poisson mixture distribution (fPMD), can be approximated by the Skellam function [22, 23]. The fPMD models the mean as a finite discrete distribution, which thus takes into account the overdispersion of RNA-seq data. The estimation of the Skellam function within a mixture model context has proved to be difficult, but we have implemented the generalized EM algorithm, integrated with the Newton-Raphson algorithm, for parameter estimation. Our algorithm allows the phenotypic plasticity of gene expression measured by RNA-seq to be estimated and tested in a quantitative manner. The new model is the first of its kind in RNA-seq data modeling, which has made it feasible to characterize the transcriptomic alterations of gene function in regulating phenotypic plasticity to environmental signals.

The Skellam model derived was used to analyze a real data set from the pharmacogentic study of breast cancer, leading to the identification of distinct gene differentiation in response to tamoxifen-sensitive and tamoxifen-resistant cells and, also, validating the practical utilization of the model. The statistical properties of the model were examined through simulation studies, with results that help geneticists determine necessary conditions for efficient and effective studies of genotype-environment interactions [11].

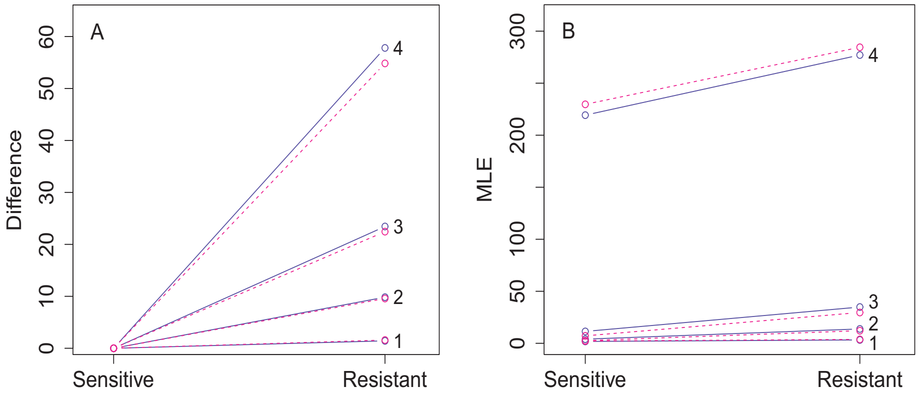

Fig. (5). Four patterns of environment-dependent gene expression (solid lines) for the simulated data estimated by the Skellam model, in a comparison with true patterns (dash lines). (**A**) Differences of gene expression for each cluster between two treatments are shown. (**B**) Actual expression values of four clusters are plotted over the treatments.

The model can be extended to construct a genetic network for the phenotypic plasticity of gene expression to an environmental stimulus [33]. This network allows a comprehensive evaluation and inference of the role of genes in mediating the phenotypic plasticity of phenotypic traits and diseases. Also, if a well-designed segregating population is available, with a set of DNA polymorphic markers genotyped through the genome and transcriptional profiles measured across different treatments, the model can be integrated with a genetic mapping approach [34] through the Skellam function to locate the expression quantitative trait loci (eQTLs) that contribute to phenotypic plasticity [35]. Although there are a few studies on the genetic mapping of plastic response for complex traits [31], eQTL mapping on phenotypic plasticity will certainly show its unique power to synthesize genomics and ecology into a unified science aimed to reveal the secrets behind life in nature.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was partially supported by grants NSF/IOS-0923975, NIH/UL1RR0330184, NIH/P50-DA10075, NIH/R01CA168676, and National Natural Science Foundation of China (11028103). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

APPENDIX

In what follows, we provide a procedure for integrating the EM and Newton-Raphson algorithms to estimate parameters contained in the likelihood (1). Define the posterior probability at which gene i belongs to cluster j as

$$\Omega_{j|i} = \frac{\pi_j p_j(m_i)}{\pi_1 p_1(m_i) + \ldots + \pi_J p_J(m_i)}. \qquad \textbf{(A1)}$$

To maximize the likelihood (1), we differentiate it with respect of unknown parameters Θ:

$$\frac{\partial \log L(\Theta \mid m)}{\partial \Theta} = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J}\Omega_{j|i}\log\pi_j + \sum_{j=1}^{J}\Omega_{j|i}\log p_j(m_i)\right\} = \sum_{i=1}^{n}\sum_{j=1}^{J}\Omega_{j|i}\log\pi_j$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{J}\Omega_{j|i}\left\{-(\lambda_{1j}+\lambda_{2j}) + \frac{m_i}{2}(\log\lambda_{1j} - \log\lambda_{2j}) + \log I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})\right\}$$

where the first and second terms are the derivatives with respect to proportion π_j and λ_{1j} and λ_{2j}, respectively. By letting the derivative with respect to π_j equal zero, we derivate a formula to estimate the mixture proportion as

$$\pi_j = \frac{\sum_{i=1}^{n}\Omega_{j|i}}{n} \qquad \textbf{(A2)}$$

Let

$$Q^* = \sum_{i=1}^{n}\Omega_{j|i}\left\{-(\lambda_{1j}+\lambda_{2j}) + \frac{m_i}{2}(\log\lambda_{1j} - \log\lambda_{2j}) + \log I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})\right\}$$

The partial derivatives of Q^* with respect to λ_{1j} and λ_{2j} are expressed as

$$\frac{\partial Q^*}{\partial \lambda_{1j}} = \sum_{i=1}^{n}\Omega_{j|i}\left\{-1 + \frac{m_i}{2}\frac{1}{\lambda_{1j}} + \sqrt{\frac{\lambda_{2j}}{\lambda_{1j}}}\frac{I'_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}{I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}\right\}$$

$$\frac{\partial Q^*}{\partial \lambda_{2j}} = \sum_{i=1}^{n}\Omega_{j|i}\left\{-1 - \frac{m_i}{2}\frac{1}{\lambda_{2j}} + \sqrt{\frac{\lambda_{1j}}{\lambda_{2j}}}\frac{I'_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}{I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}\right\}$$

Letting these derivatives equal zero, we obtain the difference of mean expression between two treatments as

$$\lambda_{1j} - \lambda_{2j} = \frac{\sum_{i=1}^{m}\Omega_{j|i}m_i}{\sum_{i=1}^{m}\Omega_{j|i}}, \qquad \textbf{(A3)}$$

if the term $\dfrac{I'_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}{I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})}$ is canceled out.

Let $\dfrac{\sum_{i=1}^{m}\Omega_{j|i}m_i}{\sum_{i=1}^{m}\Omega_{j|i}} = C_j$. Then we express λ_{2j} as a function of

λ_{1j}, i.e., $\lambda_{2j}(\lambda_{1j}) = \lambda_{2j} - C_j$ and $\lambda'_{2j}(\lambda_{1j}) = 1$. We now implement the Newton-Raphson algorithm to estimate the MLEs of λ_{1j} and λ_{2j}.

The first and second partial derivatives of Q^* with respect to λ_{1j} are expressed as

$$\frac{\partial Q^*}{\partial \lambda_{1j}} = \sum_{i=1}^{n} \Omega_{j|i} \left\{ -2 + \frac{m_i}{2}\left(\frac{1}{\lambda_{1j}} - \frac{1}{\lambda_{2j}}\right) + \frac{I'}{I}\frac{(\lambda_{1j}+\lambda_{2j})}{\sqrt{\lambda_{1j}\lambda_{2j}}} \right\}$$

$$\frac{\partial^2 Q^*}{\partial \lambda_{1j}^2} = \sum_{i=1}^{n} \Omega_{j|i} \left\{ \frac{m_i}{2}\left(-\frac{1}{\lambda_{1j}^2} + \frac{1}{\lambda_{2j}^2}\right) + \frac{(I''I - I'^2)}{I^2}\frac{(\lambda_{1j}+\lambda_{2j})^2}{\lambda_{1j}\lambda_{2j}} \right.$$
$$\left. + \frac{I'}{I}\frac{\{2\lambda_{1j}\lambda_{2j} - \frac{1}{2}(\lambda_{1j}+\lambda_{2j})^2}{(\lambda_{1j}\lambda_{2j})^{\frac{3}{2}}} \right\}$$

where

$$I = I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}})$$

$$I' = I'_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}}) = \frac{1}{2}\left\{ I_{|m_i|-1}(2\sqrt{\lambda_{1j}\lambda_{2j}}) + I_{|m_i|+1}(2\sqrt{\lambda_{1j}\lambda_{2j}}) \right\}$$

$$I'' = I''_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}}) = \frac{1}{4}\left\{ I_{|m_i|-2}(2\sqrt{\lambda_{1j}\lambda_{2j}}) + 2I_{|m_i|}(2\sqrt{\lambda_{1j}\lambda_{2j}}) \right.$$
$$\left. + I_{|m_i|+2}(2\sqrt{\lambda_{1j}\lambda_{2j}}) \right\}$$

In the Newton-Raphson iteration t, we have

$$\lambda_{1j}^{(t+1)} = \lambda_{1j}^{(t)} - \frac{\partial Q^* / \partial \lambda_{1j}}{\partial^2 Q^* / \partial \lambda_{1j}^2} \qquad \textbf{(A4)}$$

We are now able to construct a loop of the EM algorithm. In the E step, we calculate the posterior probability for gene i using (A1). In the M step, we estimate the mixture proportion of cluster j using (A2), the difference of mean expression for cluster j between the two treatments using (A3) and the mean expression of cluster j in the first treatment using (A4). The estimation of the mean expression of cluster j in the first treatment is implemented by the Newton-Raphson algorithm. The iteration is repeated until we obtain convergent estimates, which are regarded as the MLEs of the parameters.

In programming the algorithm, we found that the Newton-Raphson iteration converges extremely fast and also the rate of its convergence becomes faster and faster with the convergence of the EM iteration. Note that when the numerical solution of $I_v(x)$ occurs we can implement the saddlepoint approximation. Let $\tilde{f}(v;\lambda_1,\lambda_2)$ denote the approximation of the probability mass function $f(v;\lambda_1,\lambda_2)$. Then we have

$$I_v(x) \approx e^{-x}\tilde{f}\left(v;\frac{x}{2},\frac{x}{2}\right).$$

REFERENCES

[1] Schlichting, C.D. Phenotypic integration and environmental change. *BioScience,* **1989**, *39*, 460–464.

[2] Sultan, S.E. Phenotypic plasticity for plant development, function and life history. *Trends Plant Sci.,* **2000,** *5*, 537–542.

[3] Beldade, P.; Mateus, A.R.; Keller, R.A. Evolution and molecular mechanisms of adaptive developmental plasticity. *Mol. Ecol.,* **2011**, *20*, 1347-1363.

[4] Sommer, R.J.; Ogawa, A. Hormone signaling and phenotypic plasticity in nematode development and evolution. *Curr. Biol.,* **2011**, *21*, R758-R766.

[5] Bateson, P.; Barker, D.; Clutton-Brock, T. Developmental plasticity and human health. *Nature,* **2004**, *432*, 419-421.

[6] Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature,* **2007**, *447*, 433-439.

[7] Hochberg, Z.; Feil, R.; Constancia, M.. Child health, developmental plasticity, and epigenetic programming. *Endocrine Rev.,* **2011**, *32*, 159-224.

[8] Burdge, G.C.; Lillycrop, K.A. Nutrition, epigenetics, and developmental plasticity: implications for understanding human disease. *Ann. Rev. Nutrit.,* **2011**, *30*, 315–339.

[9] Schlichting, C.D.; Smith, H. Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. *Evol. Ecol.,* **2002**, *16*, 189-211.

[10] Li, Y.; Álvarez, O.A.; Gutteling, E.W. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.,* **2006**, *2*, e222.

[11] Gibson, G. The environmental contribution to gene expression profiles. *Nat. Rev. Genet.,* **2008**, *9*, 575–581.

[12] Smith, E.N.; Kruglyak, L. Gene-environment interaction in yeast gene expression. *PLoS Biol.,* **2008**, *6*, e83.

[13] Mortazavi, A. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods,* **2008**, *5*, 621–628.

[14] Huang, W.; Umbach, D.M.; Vincent-Jordan, N. Efficiently identifying genome-wide changes with next-generation sequencing data. *Nucleic Acids Res.,* **2011**, *39*, e130.

[15] Wang, Y.Q.; Xu, M.; Wang, Z. How to cluster gene expression dynamics in response to environmental signals. *Brief. Bioinform.,* **2012**, *13*, 162–174.

[16] Ma, L.; Li, J.; Qu, L. Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways. *Plant Cell,* **2001**, *13*, 2589-2607.

[17] Tepperman, J.M.; Zhu, T.; Chang, H.S. Multiple transcription factor genes are early targets of phytochrome A signaling. *Proc. Nat. Acad. Sci. U S* A , **2001**, *98*, 9437-9442.

[18] Gao, D.; Kim, J.; Kim, H. A survey of statistical software for analysing RNA-seq data. *Hum. Genomics,* **2010**, *5*, 56-60.

[19] Wang, N.; Wang, Y.; Hao, H. A bi-Poisson model for clustering gene expression profiles by RNA-seq. *Brief Bioinform.,* **2013,** *15*(4), 534-41.

[20] Jiang, L.B.; Mao, K.; Wu, R.L. A skellam model to identify differential patterns of gene expression induced by environmental signals. *BMC Genom*, **2014**, *15*, 772.

[21] Abramowitz, M.; Stegun, I.A. (Eds.) Modified Bessel functions I and K. Sections 9.6–9.7 in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing, pp. 374–378. New York: Dover, **1972**.

[22] Skellam, J.G. The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. Roy. Stat. Soc. Ser. A.,* **1946**, *109*, 296.

[23] Karlis, D.; Ntzoufras, I. Analysis of sports data using bivariate Poisson models. *J. Roy. Stat. Soc. Ser. D.,* **2003**, *52*, 381-393.

[24] Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25.

[25] Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106.

[26] Glaus, P.; Honkela, A.; Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bionformatics,* **2012**, *28*, 17-21.

[27] Jaiyesimi, I.A.; Buzdar, A.U.; Decker, D.A. Use of tamoxifen for breast cancer: twenty-eight years later. *J. Clin. Oncol.,* **1995**, *13*, 513-529.

[28] EBCTCG. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet,* **2005**, *365*, 1687-1717.

[29] Huber-Keener, K.J.; Liu, X.; Wang, Z. Differential patterns of mRNA transcriptomes in response to tamoxifen-sensitive and tamoxifen-resistant cells for breast cancer. *PLoS ONE,* **2012**, *7*(7), e41333.

[30] Ward, J.H. Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.,* **1983**, *58*, 236-244.

[31] Wang, Z.; Pang, X.; Lv, Y. A dynamic framework for quantifying the genetic architecture of phenotypic plasticity. *Brief. Bioinform.,*

2013, *14*, 82-95.

[32] Wu, R.L. The detection of plasticity genes in heterogeneous environments. *Evolution,* **1998**, *52*, 967–977.

[33] D'haeselee, P.; Liang, S.; Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics,* **2000**, *16*, 707-725.

[34] Wu, R.L.; Ma, C.X.; Casellam G. Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL. Springer-Verlag, New York, **2007**.

[35] Berg, A.; Li, N.; Tong, C. Functional mapping of expression quantitative trait loci that regulate oscillatory gene expression. *Methods Mol. Biol*., **2011**, *734*, 241-255.