A bi-Poisson model for clustering gene expression profiles by RNA-seq

Ningtao Wang*, Yaqun Wang*, Han Hao, Luojun Wang, Zhong Wang, Jianxin Wang and Rongling Wu

Submitted: 29th January 2013; Received (in revised form): 31st March 2013

Abstract
With the availability of gene expression data by RNA-seq, powerful statistical approaches for grouping similar gene expression profiles across different environments have become increasingly important. We describe and assess a computational model for clustering genes into distinct groups based on the pattern of gene expression in response to changing environment. The model capitalizes on the Poisson distribution to capture the count property of RNA-seq data. A two-stage hierarchical expectation-maximization (EM) algorithm is implemented to estimate an optimal number of groups and mean expression amounts of each group across two environments. A procedure is formulated to test whether and how a given group shows a plastic response to environmental changes. The impact of gene-environment interactions on the phenotypic plasticity of the organism can also be visualized and characterized. The model was used to analyse an RNA-seq dataset measured from two cell lines of breast cancer that respond differently to an anti-cancer drug, from which genes associated with the resistance and sensitivity of the cell lines are identified. We performed simulation studies to validate the statistical behaviour of the model. The model provides a useful tool for clustering gene expression data by RNA-seq, facilitating our understanding of gene functions and networks.

Keywords: RNA-seq; Poisson distribution; EM algorithm; breast cancer cell lines

INTRODUCTION
RNA-seq, a next-generation sequencing technique for measuring messenger RNA (mRNA) levels for all genes, had been revolutionizing our tools to study genome structure, gene expression and genetic networks [1–4]. In nature, the environment often changes rapidly and stochastically, during which a living creature would alter the level and pattern of its gene expression to efficiently accommodate this change [5]. The power of RNA-seq is to detect and identify specific genes that are associated with the environment-induced response of the organism. Thus, a particular experiment using RNA-seq contains two or more treatment groups that vary in environmental, developmental or physiological factors, allowing the pattern of gene differential expression to be tested [6, 7]. More recently, RNA-seq experiments have examined multiple treatment groups to elucidate reaction norms of gene expression profiles and understand the transcriptional network.
associated with the development of a phenotypic trait [8].

To detect the environment-induced change of gene expression, cluster analysis has been widely used as a computational tool that can separate genes into different groups based on their expression patterns, where genes within each group tend to be functionally related [9–14]. However, most existing approaches for cluster analysis have not accommodated the particular properties of RNA-seq data. First, the level of gene expression by RNA-seq is described by the enumeration of short sequences, or reads, mapped to each gene, which is defined as a collection of exons [15]. As a type of count data, three discrete probability distributions, binomial, Poisson and negative binomial (NB), have been used as a computational tool that can separate gene expression, cluster analysis has been widely used to identify genes that belong to a mixture model framework in which each mixture is represented by a distinct pattern of expression profile. The mean levels of gene expression for a mixture model is described by a joint probability density function.

In this article, we describe and validate a bi-Poisson model for clustering genes based on their differential expression over two different environments. The model implements the bi-Poisson distribution within a mixture model framework in which each mixture is represented by a distinct pattern of expression profile. The mean levels of gene expression for a cluster in different environments can be calculated through the two-stage hierarchical expectation–maximization (EM) algorithm. By comparing environment-induced differences of gene expression, we can characterize the genetic machinery of the response of the organism to environmental changes. The model was used to analyse an RNA-seq data collected for two contrast cell lines, one sensitive to an anti-cancer drug and the other resistant to the same drug [17]. We performed computer simulation to investigate the statistical behavior of the model, validating its usefulness and utilization in practice.

MODEL
Mixture model-based likelihood
Suppose we initiate a transcriptomic study in which the organism is measured for reads of \( n \) genes expressed in two different treatments, such as different tissues, different stages of development or different temperatures [6]. The expression reads of gene \( i \) measured in the two treatments 1 and 2 are denoted as \( m_{i1} \) and \( m_{i2} \), respectively. By comparing these two values, the extent to which this gene is differentially expressed across the treatment can be characterized. Thus, if cluster analysis is based simultaneously on two values of expression in the two treatments, we can discern different groups of genes per their functional similarities and differences in plastic response to the change of environment.

For any gene \( i \), it should arise from one (and only one) of the \( J \) groups that are classified on the basis of two expression values in different treatments. The joint likelihood of the expression data \( \mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2) \) of \( n \) genes is written as

\[
L(\Theta | \mathbf{m}) = \prod_{i=1}^{n} \left[ \pi_1 H_1 (m_{i1}, m_{i2}) + \ldots + \pi_J H_J (m_{i1}, m_{i2}) \right],
\]

(1)

where \( \Theta \) is a set of unknown parameters; \( \pi_i \) is the probability of group \( j \) \((j = 1, \ldots, J)\) in the total genes; and \( H_j (m_{i1}, m_{i2}) \) is the density function of two expression values for gene \( i \) that belongs to group \( j \) in the two treatments. In this study, we use a bivariate Poisson distribution function to describe \( H_j (m_{i1}, m_{i2}) \), which is specified by the mean values of gene expression in treatment 1 \((\lambda_{j1})\) and 2 \((\lambda_{j2})\) and the covariance of gene expression values between the two treatments \((\lambda_{j})\).

Let \( Z_1 \), \( Z_2 \) and \( Z \) denote three independent random Poisson variables with mean \( \lambda_{j1} \), \( \lambda_{j2} \) and \( \lambda_j \) for group \( j \), respectively. The three variables are expressed as two dependent random variables \( M_1 = Z_1 + Z_2 \) and \( M_2 = Z_2 + Z \). A bivariate Poisson distribution of \( M_1 \) and \( M_2 \) for gene \( i \) is described by a joint probability density function, expressed as

\[
f_i (M_{i1} = m_{i1}, M_{i2} = m_{i2} | \Lambda_i) = \exp \left[ - (\lambda_{j1} + \lambda_{j2} + \lambda_j) \right] \frac{\lambda_{j1}^{m_{i1}} \lambda_{j2}^{m_{i2}} \exp (-\lambda_{j1} - \lambda_{j2})}{m_{i1}! m_{i2}!} \sum_{l=0}^{\min(m_{i1}, m_{i2})} \left( \begin{array}{c} m_{i1} \\ l \end{array} \right) \left( \begin{array}{c} m_{i2} \\ l \end{array} \right) \left( \frac{\lambda_{j1} \lambda_{j2}}{\lambda_j} \right)^l \left( \frac{\lambda_j}{\lambda_{j1} \lambda_{j2}} \right),
\]

(2)

where \( \lambda_{j1} \) and \( \lambda_{j2} \) are the mean expression values of all genes that belong to group \( j \) in treatment 1 and 2, respectively, and \( \lambda_j \) is the covariance of group \( j \) between the two treatments, with the three parameters arrayed in \( \Lambda_i = (\lambda_{j1}, \lambda_{j2}, \lambda_j) \). Here, \( H_j (m_{i1}, m_{i2}) \) in the mixture model (1) is specified by \( f_i (M_{i1} = m_{i1}, M_{i2} = m_{i2} | \Lambda_i) \).
Estimation via the EM algorithm

Given that \((m_1, m_2)\) are bivariate Poisson variables, we define \((m_1 - Z_n, m_2 - Z_n, Z_n)\) as three independent Poisson variables, where \(Z_n\) is a latent variable and can be estimated by an EM algorithm. Different from a general bi-Poisson model, the likelihood of \((m_1, m_2)\) is formulated within a mixture-model framework \((1)\) whose estimation is based on the implementation of the EM algorithm. Thus, we implement a two-stage hierarchal EM algorithm to estimate the parameters \(\Lambda_j\) of the likelihood \((1)\).

In the E step, we calculate the conditional expectation of \(Z_n\) by

\[
Z_i^{(l)} = E[Z_i | m_1, m_2, \pi^{(l-1)}, \Lambda^{(l-1)}] = \sum_{z=0}^{\min(m_1, m_2)} \pi_i^{(l-1)} \frac{H^*_i(m_1, m_2, z | \Lambda^{(l-1)})}{\sum_{z=0}^{\min(m_1, m_2)} \pi_i^{(l-1)} \frac{H^*_i(m_1, m_2, z | \Lambda^{(l-1)})}{}}
\]

where \(H^*_i\) is the density of joint distribution of \((m_1, m_2, Z_n)\). Meanwhile, we calculate the posterior probability of gene \(i\) that belongs to group \(j\),

\[
\Omega_{ij}^{(l)} = \frac{\pi_i^{(l-1)} H^*_i(m_1, m_2 | \Lambda^{(l-1)})}{\sum_{j=1}^{n} \pi_i^{(l-1)} H^*_i(m_1, m_2 | \Lambda^{(l-1)})}
\]

In the M step, we obtained the estimates of parameters \(\pi_i\) and \(\Lambda_j\) by using

\[
\pi_i^{(l)} = \frac{n}{\sum_{i=1}^{n} \Omega_{ij}^{(l)}}
\]

\[
\lambda_{ij}^{(l)} = \frac{\sum_{i=1}^{n} \Omega_{ij}^{(l)} (m_1 - Z_i^{(l)})}{\sum_{i=1}^{n} \Omega_{ij}^{(l)}}
\]

\[
\lambda_{j1}^{(l)} = \frac{\sum_{i=1}^{n} \Omega_{ij}^{(l)} (m_2 - Z_i^{(l)})}{\sum_{i=1}^{n} \Omega_{ij}^{(l)}}
\]

\[
\lambda_{j2}^{(l)} = \frac{\sum_{i=1}^{n} \Omega_{ij}^{(l)} Z_i^{(l)}}{\sum_{i=1}^{n} \Omega_{ij}^{(l)}}
\]

The E and M steps are iterated between equations \((3-8)\) until the estimates of the unknown parameters converge to stable values. The estimates obtained this way are the maximum likelihood estimates (MLEs) of the parameters.

Hypothesis tests

The determination of an optimal number of components in the mixture model \((1)\) is a challenging issue in statistics. Commonly used model selection criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), can be used to estimate the number of gene groups for a particular RNA-seq dataset. By increasing the number of groups for the model successively starting with two, these AIC or BIC values are calculated. The optimal number of groups is one that provides a minimum AIC or BIC value. At each time for a given group number, we pose one more restriction on the parameter space of across-treatment differences for any two groups \(j, l\) as

\[
\lambda_{j2} - \lambda_{l1} \leq \lambda_{i2} - \lambda_{i1}, \quad \forall j < l = 1, \ldots, J
\]

in order to make the mixture model identifiable.

After an optimal number of gene clusters is determined, three biologically meaningful tests can be formulated as follows:

(i) For a given group, we hope to know whether its genes are differently expressed between the two treatments. This can be tested by testing

\[
H_0 : \lambda_{j1} = \lambda_{j2} \text{ vs. } H_1 : \lambda_{j1} \neq \lambda_{j2}, \quad \forall j = 1, \ldots, J
\]

If the \(H_0\) is accepted, this means that group of genes is stable across the treatments. Otherwise, they display different amounts of expression between the two treatments, in which case they can be used as a predictor of environment-induced changes.

(ii) For a pair of groups \(j\) and \(l\), we hope to know whether they interact with each other to determine environment-induced changes. This can be done by testing:

\[
H_0 : \lambda_{j1} - \lambda_{l1} = \lambda_{j2} - \lambda_{l2} \text{ vs. } H_1 : \lambda_{j1} - \lambda_{l1} \neq \lambda_{j2} - \lambda_{l2}, \quad \forall j < l = 1, \ldots, J.
\]

If the \(H_0\) is rejected, this means that these two groups of genes have significant interaction effects on biological changes between the treatments.

(iii) For a pair of groups \(j\) and \(l\), we hope to know whether they have the same across-environment covariance of gene expression. This can be done by testing

\[
H_0 : \lambda_j = \lambda_l \text{ vs. } H_1 : \lambda_j \neq \lambda_l, \quad \forall j < l = 1, \ldots, J.
\]
If the $H_0$ is accepted, this means that these two groups of genes have the same covariance of gene expression between the treatments.

For each of the hypotheses (9–11), we calculate the log-likelihood ratio (LR) test statistics, which are thought of being chi-square distributed with the degree of freedom equaling the difference of the number of parameters to be estimated under the $H_0$ and $H_1$. If these tests are incorporated by a particular environmental signal, we can better understand the relationship between gene expression and environment.

**MODEL VALIDATION**

**Worked example**

By mediating the oestrogen receptor (ER) to prevent ER-mediated transcription, tamoxifen can block the effects of oestrogen in breast cancer cells and, therefore, is used in the adjuvant hormonal therapy of breast cancer patients. There have been some examples of success in treating ER-negative breast tumours by tamoxifen, but its efficacy is often affected by drug resistance. In a recent study, aimed to reveal the global mechanisms of gene expression and signalling pathway alterations for tamoxifen resistance, by Huber-Keener et al. [17], the transcriptomes of breast cancer cells were measured for two cell lines, one being tamoxifen-sensitive and the second being tamoxifen-resistant, for a total of 23 561 mRNA genes using RNA-seq.

The model developed was used to analyse differences of these RNA genes expressed in tamoxifen-sensitive and tamoxifen-resistant cell lines. Because it incorporates the information of sample size, we used BIC as the model-selection criterion. The BIC values calculated suggest that an optimal number of gene groups is 10 for this dataset (Figure 1). For each group $j$, the mean values of gene expression ($\lambda_{j1}$ and $\lambda_{j2}$) in two cell types, as well as the covariance of gene expression values between the two lines ($\lambda_j$), were estimated, with reasonable good standard errors (Table 1). In practical calculations, the estimate of $\lambda_j$ is sensitive to the choice of initial values. To obtain a global maxima, multiple initial values need to be selected and compared. Figure 2 illustrates the gene expression plot of 10 clusters detected over two cell lines. It can be seen that all these 10 clusters not only display different amounts of gene expression (Figure 2A), but also vary dramatically in the change of expression amount from sensitive to resistant lines (Figure 2B). Hypothesis test (9) provides information about the significance of expression difference between two cell lines. Of these 10 clusters, gene expression of clusters 1 and 2 tends to be stable from sensitive to resistant lines, although the change of gene expression is significant ($P<0.05$). All others are strikingly up-regulated, but to different extents. The amounts of up-regulation are the highest for clusters 9 and 10, followed by cluster 8, clusters 5, 6 and 7 and clusters 3 and 4. In general, cell line-specific changes of gene expression are positively associated with the amounts of expression in both cell lines.

**Figure 1:** Plot of BIC values over the number of clusters calculated from a transcriptomic study involving two breast cancer cell lines that are sensitive and resistant to tamoxifen, respectively.

**Table 1:** Maximum likelihood estimates of mean expression values of genes ($\lambda_{j1}$ and $\lambda_{j2}$, $j = 1, \ldots, 10$) for 10 distinct groups in two breast cancer cell lines that are sensitive and resistant to tamoxifen, respectively.

<table>
<thead>
<tr>
<th>Group</th>
<th>Proportion</th>
<th>$\lambda_{j1}$</th>
<th>$\lambda_{j2}$</th>
<th>$\lambda_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.204 (0.016)</td>
<td>0.79 (0.078)</td>
<td>1.61 (0.121)</td>
<td>1.9e-3 (4.0e-4)</td>
</tr>
<tr>
<td>2</td>
<td>0.185 (0.018)</td>
<td>2.91 (0.252)</td>
<td>5.78 (0.336)</td>
<td>0.029 (0.005)</td>
</tr>
<tr>
<td>3</td>
<td>0.205 (0.018)</td>
<td>5.40 (0.236)</td>
<td>12.8 (0.680)</td>
<td>0.123 (0.014)</td>
</tr>
<tr>
<td>4</td>
<td>0.099 (0.018)</td>
<td>10.0 (0.258)</td>
<td>21.3 (0.828)</td>
<td>0.394 (0.051)</td>
</tr>
<tr>
<td>5</td>
<td>0.016 (0.013)</td>
<td>11.5 (0.700)</td>
<td>34.4 (1.051)</td>
<td>0.737 (0.062)</td>
</tr>
<tr>
<td>6</td>
<td>0.066 (0.012)</td>
<td>20.2 (1.093)</td>
<td>44.9 (1.564)</td>
<td>1.647 (0.113)</td>
</tr>
<tr>
<td>7</td>
<td>0.022 (0.008)</td>
<td>31.8 (3.624)</td>
<td>62.7 (4.501)</td>
<td>3.364 (0.378)</td>
</tr>
<tr>
<td>8</td>
<td>0.037 (0.018)</td>
<td>18.4 (3.026)</td>
<td>68.2 (4.364)</td>
<td>2.210 (0.440)</td>
</tr>
<tr>
<td>9</td>
<td>0.056 (0.009)</td>
<td>34.1 (1.285)</td>
<td>103 (4.162)</td>
<td>5.529 (0.423)</td>
</tr>
<tr>
<td>10</td>
<td>0.016 (0.004)</td>
<td>62.9 (4.025)</td>
<td>135 (7.705)</td>
<td>12.54 (1.894)</td>
</tr>
</tbody>
</table>

The covariance of gene expression between two cell lines ($\lambda_j$) is also given. The standard errors (in parentheses) of the estimates are calculated from 1000 bootstrapping samples.
A test was made to investigate whether a particular pair of gene groups interacts with the environment to determine the difference of cell lines in response to tamoxifen. Table 2 lists the test of significance for such gene–gene interactions. Except for a few pairs, most pairs of gene groups exhibit significant gene–environment interactions.

Computer simulation
Simulation studies were carried out to examine the statistical behaviour of the model by investigating the precision of parameter estimation. By mimicking the tamoxifen example as described above, we simulated read data of 1000 transcript genes with 10 distinct clusters. The model was used to analyse the simulated data and find the number of clusters. For 1000 simulation replicates, the model can always correctly find 10 clusters as shown from BIC values. Table 3 tabulates the maximum likelihood estimates of $\lambda_1$, $\lambda_2$ and $\lambda_3$ for different groups, in a comparison with their true values. In general, the mean values and covariance of gene expression between different treatments can be reasonably well estimated. As seen from Figure 3, the estimated curves of gene expression for each group are broadly consistent with the true curves, suggesting the accuracy of estimation and test for gene differentiation across the treatment.

As one of the most commonly used clustering algorithm, we compared K-means with our mixture-based model. In our simulation condition as given above, both approaches provide a similar result. However, because the two approaches have different underlying principles, they are interpreted differently. K-means clustering tends to identify clusters of similar spatial extents, whereas our model finds different clusters based on their shape of gene expression over treatments. Thus, by providing biologically testable hypotheses (9–11), the model should be of greater value from a biological perspective.

DISCUSSION
Gene expression data by RNA-seq provide a valuable engine to study the regulatory mechanisms of the transcriptional response to a stimulus [5, 7]. Here, we describe a computational tool to address the challenge of analysing such complex data. Our bi-Poisson model is an integrative approach where the expression of genes in different environments is analysed at the same time so that every cluster captures a group of genes with a similar pattern of response to the environment. By exploiting the fact that the same group of co-regulated genes responds to the change of environment through similar machinery, a set of prototypical responses can be

Table 2: Hypothesis tests for gene–environment interactions for different pairs of gene group

<table>
<thead>
<tr>
<th>Group Pair</th>
<th>Test Static</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 versus 2</td>
<td>60.54</td>
<td>7.22e-15</td>
</tr>
<tr>
<td>2 versus 3</td>
<td>39.29</td>
<td>3.65e-10</td>
</tr>
<tr>
<td>3 versus 4</td>
<td>16.41</td>
<td>5.09e-05</td>
</tr>
<tr>
<td>4 versus 5</td>
<td>43.24</td>
<td>4.84e-11</td>
</tr>
<tr>
<td>5 versus 6</td>
<td>2.15</td>
<td>0.143</td>
</tr>
<tr>
<td>6 versus 7</td>
<td>4.20</td>
<td>0.0403</td>
</tr>
<tr>
<td>7 versus 8</td>
<td>8.52</td>
<td>3.53e-03</td>
</tr>
<tr>
<td>8 versus 9</td>
<td>22.34</td>
<td>2.29e-06</td>
</tr>
<tr>
<td>9 versus 10</td>
<td>19.1</td>
<td>1.67</td>
</tr>
</tbody>
</table>

Figure 2: Mean values of gene expression for 10 distinct groups in sensitive and resistant cell lines of breast cancer. (A) Absolute values of gene expression in the two cell lines. (B) Differences of gene expression from sensitive to resistant cell lines.
estimated and tested in a function space. These can be used as a meaningful prior to characterize the functional relationship between genes and environment.

To demonstrate the properties of our method, we applied the model to a dataset describing the response of breast cancer cell lines to an anticancer drug, tamoxifen [17]. As shown, the resulting clusters capture distinct parts of the transcriptomic response that represent separate biological functions related to the sensitivity and resistance of cancer cells to the drug. Also, the estimated model parameters are biologically meaningful and allow us to provide a finer description of the differences between the two cell lines. To validate the findings from the data by the new model, we performed simulation studies, suggesting that the model can be used in practice. If the practical data contain more than two treatments, then the model should be extended to include a multivariate Poisson distribution [18, 19].

In this article, we describe a general pipeline for clustering transcriptomic response to changing environments based on a simple bi-Poisson distribution. Given a complex data with great variability, other more sophisticated distributions should be incorporated to provide a better flexibility of fit. These include the negative binomial distribution as a natural extension of the Poisson distribution, which can account for data dispersion [20], and the generalized Poisson distribution [21], which can model overdispersion. In general, clustering genes of differential expression is not the final step of the analysis.

Table 3: Results of parameter estimates from simulated data by mimicking the data structure of breast cancer transcriptomic study

<table>
<thead>
<tr>
<th>Group</th>
<th>Proportion</th>
<th>$\lambda_{ij}$</th>
<th>$\lambda_{kj}$</th>
<th>$\lambda_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>MLE</td>
<td>True</td>
<td>MLE</td>
</tr>
<tr>
<td>1</td>
<td>0.206</td>
<td>0.206 (0.008)</td>
<td>0.80</td>
<td>0.80 (0.075)</td>
</tr>
<tr>
<td>2</td>
<td>0.192</td>
<td>0.192 (0.011)</td>
<td>2.96</td>
<td>2.97 (0.169)</td>
</tr>
<tr>
<td>3</td>
<td>0.204</td>
<td>0.204 (0.011)</td>
<td>5.45</td>
<td>5.45 (0.216)</td>
</tr>
<tr>
<td>4</td>
<td>0.095</td>
<td>0.095 (0.010)</td>
<td>10.2</td>
<td>10.1 (0.468)</td>
</tr>
<tr>
<td>5</td>
<td>0.104</td>
<td>0.104 (0.009)</td>
<td>11.4</td>
<td>11.4 (0.426)</td>
</tr>
<tr>
<td>6</td>
<td>0.067</td>
<td>0.067 (0.006)</td>
<td>19.9</td>
<td>20.0 (0.740)</td>
</tr>
<tr>
<td>7</td>
<td>0.020</td>
<td>0.020 (0.003)</td>
<td>32.3</td>
<td>32.2 (1.708)</td>
</tr>
<tr>
<td>8</td>
<td>0.037</td>
<td>0.037 (0.004)</td>
<td>17.7</td>
<td>17.7 (0.847)</td>
</tr>
<tr>
<td>9</td>
<td>0.058</td>
<td>0.058 (0.001)</td>
<td>33.9</td>
<td>33.9 (0.747)</td>
</tr>
<tr>
<td>10</td>
<td>0.017</td>
<td>0.017 (3.8e-4)</td>
<td>62.5</td>
<td>62.5 (1.791)</td>
</tr>
</tbody>
</table>

The MLE from the model are compared with the true values for each parameter. The standard deviations of the MLEs (in parentheses) are calculated from 1000 simulation replicates.

Figure 3: Comparison of estimated gene expression (solid) with true values (broken) for 10 distinct groups from simulated data by mimicking the transcriptomic study of breast cancer. (A) Absolute values of gene expression in the two cell lines. (B) Differences of gene expression from sensitive to resistant cell lines.
Other analyses, such as gene set testing, network inference and knowledge databases, should be followed [22–24]. A comprehensive model of integrating gene clustering and these follow-up analyses is worthy to derive, enabling geneticists to extract biological insight from their gene expression data.

The final goal of our genetic analysis is to draw a genotype–phenotype map by establishing a more complete picture of gene regulation [25]. In so doing, there is a pressing need to integrate the results of RNA-seq data with other sources of biological data [26]. In conjunction with genotyping data, RNA-seq can be used to identify expression quantitative trait loci or eQTLs that are responsible for variation in gene expression [27, 28]. Several studies have integrated expression data with transcription factor binding, RNA interference, histone modification and DNA methylation information, aimed to better understand the regulatory mechanisms of gene expression [29–31]. In addition, most studies of gene expression by RNA-seq are still performed in a static state, but increasing recognition has been given to the role of dynamic gene expression in constructing regulatory networks [8–14]. To model dynamic changes of gene expression in response to environmental stimuli, more advanced statistical model such as longitudinal data analysis integrating the multivariate Poisson distribution [18, 19] is required, a topic deserving further investigation.

Key points

- One of the fundamental mechanisms for an organism to respond to environmental fluctuations is to alter its gene expression.
- By grouping gene expression profiles across environments, cluster analysis can shed light on gene functions and regulation involved in a biological process.
- RNA-seq has provided an unprecedented opportunity to study global gene expression by measuring mRNA levels for all genes.
- We described and assessed a clustering algorithm for the modeling and analysis of RNA-seq data based on a bi-Poisson distribution of gene reads measured under two conditions.

Acknowledgements

We thank Prof. J.-M. Yang for providing the published RNA-seq data to test and validate our model.

FUNDING

National Center for Advancing Translational Sciences (NCATS) (UL1 TR000127). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References


