BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 197–205
doi:10.1093/bioinformatics/btt632

Gene expression

Advance Access publication November 4, 2013

Model-based clustering for RNA-seq data

Yaqing Si[1,2,*], Peng Liu[2,*], Pinghua Li[3] and Thomas P. Brutnell[4]

[1]School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China, [2]Department of Statistics, Iowa State University, Ames, IA 50011, USA, [3]Institute of Tropical Biosciences and Biotechnology (ITBB), Chinese Academy of Tropical Agriculture Sciences (CATAS), Haikou, Hainan 571101, China and [4]Enterprise Institute for Renewable Fuels, Donald Danforth Plant Science Center, St. Louis, MO 63132, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: RNA-seq technology has been widely adopted as an attractive alternative to microarray-based methods to study global gene expression. However, robust statistical tools to analyze these complex datasets are still lacking. By grouping genes with similar expression profiles across treatments, cluster analysis provides insight into gene functions and networks, and hence is an important technique for RNA-seq data analysis.

Results: In this manuscript, we derive clustering algorithms based on appropriate probability models for RNA-seq data. An expectation-maximization algorithm and another two stochastic versions of expectation-maximization algorithms are described. In addition, a strategy for initialization based on likelihood is proposed to improve the clustering algorithms. Moreover, we present a model-based hybrid-hierarchical clustering method to generate a tree structure that allows visualization of relationships among clusters as well as flexibility of choosing the number of clusters. Results from both simulation studies and analysis of a maize RNA-seq dataset show that our proposed methods provide better clustering results than alternative methods such as the K-means algorithm and hierarchical clustering methods that are not based on probability models.

Availability and implementation: An R package, MBCluster.Seq, has been developed to implement our proposed algorithms. This R package provides fast computation and is publicly available at http://www.r-project.org.

Contact: sy@swufe.edu.cn; pliu@iastate.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2013; revised on July 29, 2013; accepted on October 29, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have revolutionized studies of genome structure, gene expression and epigenetics (Metzker, 2010; Wang *et al*., 2010). One important application of NGS technologies is in the study of gene expression by measuring messenger RNA levels for all genes in a sample. This technology is called RNA-seq, and several reviews have described this nascent technology (Marguerat *et al*., 2008; Metzker, 2010; Wang *et al*., 2010, 2009). Here we briefly describe how RNA-seq data can be generated. The complete set of messenger RNA molecules are first extracted from a sample and converted to a library of short complementary DNA fragments. Then these fragments are sequenced simultaneously by NGS technology. The resulting millions of short sequences, which are commonly called reads, are then aligned to a reference genome or reference transcripts. Gene expression is measured by the enumeration of reads mapped to each gene where the gene can be defined as a collection of exons or other appropriate definitions given the context of a study (Bullard *et al*., 2010). The resulting RNA-seq data are essentially digital signals that can be used to quantify levels of gene expression (Marguerat *et al*., 2008; Wang *et al*., 2009). This differs from microarray technologies that measure gene expression by fluorescence intensities detected from hybridized samples. Inescapable factors such as cross-hybridization, secondary structure of the DNA and technical challenges associated with fluorescent detection used in microarray analysis limit both the sensitivity and dynamic range. Compared with microarray technologies, NGS technologies permit quantitative measures of gene expression over a much larger dynamic. These advantages have rapidly accelerated the adoption of the NGS technologies in studies of gene expression and present new challenges to data analysis.

In the pioneering studies using RNA-seq, only two treatment groups were analyzed (Marioni *et al*., 2008; Sultan *et al*., 2008). More recently, RNA-seq experiments that examined multiple treatment groups have been published. For example, Li *et al*. (2010) carefully selected a developing leaf from a corn plant that captures multiple stages of photosynthetic differentiation. They exploited Illumina sequencing technologies to profile gene expression from four representative sections of the leaf blade. One major goal of this study was to survey gene expression profiles along different developmental stages to gain understanding of the transcriptional network associated with the development of C4 photosynthesis. In this endeavor, cluster analysis is an important tool as it often reveals groups of genes with similar expression patterns, where genes within such groups tend to be functionally related.

Li *et al*. (2010) took a heuristic approach by applying the K-means algorithm to partition log-transformed data for the differentially expressed genes. The K-means algorithm starts from an initial partition of the objects (genes) and proceeds by iteratively calculating the centers (means) of clusters and reassigning each object to the closest cluster according to some measurement of distance such as Euclidean distance. This iteration continues until no more reassignments take place. Although this heuristic approach is easy to implement, its

*To whom correspondence should be addressed

performance was not evaluated for RNA-seq data analysis. Studies of clustering algorithms with microarray data revealed that heuristic algorithms performed worse than model-based algorithms (Yeung *et al*., 2001). Surprisingly, there has been few published statistical research to examine cluster analysis of RNA-seq data, although it is urgently needed due to the huge amount of data being generated. Model-based algorithms for microarray data are based on finite mixture of normal distributions and cannot be directly applied to RNA-seq data that are discrete counts and often skewed. RNA-seq data have been modeled using Poisson (Bullard *et al*., 2010; Marioni *et al*., 2008) or negative binomial (NB) distributions (Robinson *et al*., 2010). Witten (2011) describes a hierarchical clustering method to cluster samples (experimental units) based on the RNA-seq data of all genes within each sample using Poisson model and dissimilarity measure based on likelihood ratio statistics. Often the case, as in Li *et al*. (2010), clustering gene expression profiles is of interest. In this article, we aim to cluster genes based on the differential expression patterns across treatments using model-based statistical methods. In other words, we are interested in grouping genes that share the same or similar expression fold-changes with respect to the mean expression level across all treatments. To do this, we derive model-based clustering algorithms for cluster genes based on either Poisson or NB models for RNA-seq data, and we evaluate the performance of the model-based approach and heuristic algorithms including the K-means method to cluster genes.

We describe the Poisson and NB distributions in Section 2 and show how our model-based clustering method handles both probability models in a unified fashion. We present an expectation-maximization (EM) algorithm for estimating the model parameters and cluster membership in Section 3.1. In addition, a model-based initialization algorithm is proposed in Section 3.2 to reduce the dependence on the initialization. We also describe two stochastic versions of EM algorithms in Section 3.3 that are intended to reduce the chance of being trapped at local solutions. A model-based hierarchical algorithm is proposed in Section 3.4 to generate a hierarchical structure of the clusters and allow more flexibility of choosing cluster numbers. In Section 4, we simulate data and compare the proposed method with others using three commonly used criteria: sensitivity, specificity and mutual information (MI) (Booth *et al*., 2008; Strehl and Ghosh, 2002; Woodard and Goldszmidt, 2011). In Section 5, we apply the model-based method to the data from Li *et al*. (2010) and evaluate our results by comparing the clusters with gene annotations. We summarize in Section 6 that our results from extensive simulation studies and an analysis of an RNA-seq dataset all show that our proposed method outperforms alternative methods, namely, the K-means algorithm and self-organizing map (SOM) (Ressom *et al*., 2003; Tamayo *et al*., 1999).

2 MODEL

Let N_{gij} denote the count of reads mapped to gene g for replicate j of treatment i for $g = 1, \cdots, G; i = 1, \cdots, I;\ j = 1, \cdots, n_i$, where G is the total number of genes of interest, I is the number of treatment groups and n_i is the number of replicates for treatment i. Two discrete probability distributions have been proposed to model RNA-seq data. The Poisson distribution has been shown to be appropriate for the RNA-seq data when only technical replicates are included (Bullard *et al*., 2010; Marioni *et al*., 2008). When there are biological replicates, RNA-seq data may exhibit more variability than expected with a Poisson distribution, i.e. the overdispersion phenomenon (Anders and Huber, 2010). The NB model proposed by Robinson and Smyth (2008) originally for serial analysis of gene expression data allows overdispersion and has been applied to RNA-seq data analysis (Anders and Huber, 2010; Robinson *et al*., 2010). We consider both distributions in this article.

2.1 Poisson distribution

Suppose N_{gij} follows a Poisson distribution with mean λ_{gij} that is parameterized as follows:

$$\log \lambda_{gij} = s_{gij} + \alpha_g + \beta_{gi} \tag{1}$$

with $\sum_{i=1}^{I} \beta_{gi} = 0$. The offset term s_{gij} is a normalization factor that may depend on the gene length and library of a sample such as the total number of mapped reads of a library. Once estimated from data, the normalization factor is often treated as known in the model (Bullard *et al*., 2010; Marioni *et al*., 2008; Robinson and Oshlack, 2010). The parameter α_g represents the geometric mean expression level of gene g across all treatments; β_{gi} measures the expression level of gene g in treatment i relative to the overall mean expression. To cluster gene expression profiles, we are interested in clustering the vectors $\beta_g = (\beta_{g1}, \cdots, \beta_{gI})$ for all G genes.

2.2 Negative binomial distribution

For the NB model, we adopt the parametrization in Robinson and Smyth (2008) by modeling the variance as

$$\mathrm{Var}(N_{gij}) = \lambda_{gij} + \phi_g \lambda_{gij}^2 \tag{2}$$

where λ_{gij} is the same as in (1) and ϕ_g is a dispersion parameter. Compared with Poisson model, an extra parameter, ϕ_g, is introduced for each gene. Robinson and Smyth (2008) described several methods to estimate ϕ_g. In this article, we estimate ϕ_g by the quasi-likelihood method. To simplify the algorithm, we treat ϕ_g as known on its estimation because our numerical studies showed this strategy produced similar clustering results to those based on the true ϕ_g values (see Section 4.3). With this strategy, the unknown parameters are the same for the Poisson and NB models, and thus we denote the likelihood function for both models by $f(\boldsymbol{N}_g|\alpha_g, \beta_g)$ for gene g where $\boldsymbol{N}_g = \{N_{gij}\}$.

3 MODEL-BASED CLUSTERING

Model-based clustering methods assume that data are generated by a mixture of probability distributions where each component corresponds to one cluster. Extensive research has been done in model-based clustering with multivariate normal mixture distributions. See, for example, Fraley and Raftery (2002) for an excellent review. In this section, we describe model-based clustering for RNA-seq data with the probability models introduced in Section 2.

The algorithms described later in the text aim to cluster gene expression profiles, which is desired in practical application.

Consequently, genes within the same cluster have similar expression profiles (denoted by β_g in our notation), but may have different overall mean expression levels (indicated by α_g). However, it is straightforward to make changes in the algorithm if the goal is to cluster according to both the overall expression levels and the expression profiles, $\alpha_g + \beta_g$.

Suppose there are K clusters and let $\mu_k = (\mu_{k1}, \cdots, \mu_{kI})$ denote the *center* of cluster k with $\sum_{i=1}^{I} \mu_{ki} = 0$ for $k = 1, \cdots, K$. The likelihood of the mixture model for gene g is $\sum_k p_k f(\boldsymbol{N}_g|\alpha_g, \beta_g = \mu_k)$, where $f(\boldsymbol{N}_g|\alpha_g, \beta_g = \mu_k)$ is the likelihood if gene g belongs to the kth cluster and p_k is the mixing proportion with $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. The likelihood function can be based on a Poisson model or NB model as described in Section 2. Taking all genes together, the likelihood is as follows:

$$L = \prod_g \sum_k p_k f(\boldsymbol{N}_g|\alpha_g, \beta_g = \mu_k) \tag{3}$$

Note that we assume independence among genes, which is likely not true in real situations. However, it is difficult, or impossible, to model and estimate the correlation among tens of thousands of genes with only several replicates and no prior knowledge about the relationship among genes. Thus, for simplicity, we take the independence assumption as in previous model-based cluster analysis for microarray studies (Yeung *et al*., 2001).

3.1 Model-based clustering with the expectation-maximization algorithm (MB-EM)

The EM algorithm has been widely applied to model-based clustering with multivariate normal mixture distributions (Fraley and Raftery, 2002). McLachlan (1997) describes an EM algorithm to fit overdispersed univariate count data in Poisson regression and logistic regression setting. Here, we derive an EM algorithm (Algorithm 1) for clustering RNA-seq gene expression profile with a mixture of Poisson or NB models. Let $Z_{gk} = 1$ if gene g belongs to the kth cluster and $Z_{gk} = 0$ otherwise. The EM algorithm views the cluster memberships $\boldsymbol{Z} = \{Z_{gk} : g = 1, \cdots, G; k = 1, \cdots, K\}$ as missing data and proceeds by iteratively calculating the conditional expectations of $\boldsymbol{Z}$ and updating the estimates for model parameters until convergence:

Algorithm 1: MB-EM Algorithm.

(i) *Initialization*: *Set* $p_k^{(1)}$ *according to prior knowledge about the cluster size. If no such information is available, let* $p_k^{(1)} = 1/K$ *for* $k = 1, \cdots, K$. *Choose* K *vectors* $\mu_1^{(1)}, \cdots, \mu_K^{(1)}$ *with* $\sum_{i=1}^{I} \mu_{ki}^{(1)} = 0$ *for* $k = 1, \cdots, K$ *as the initial set of cluster centers. See Algorithm 2 for one way to choose these* $\mu_k^{(1)}$. *Obtain the initial values of* $\alpha^{(1)} = \{\alpha_{gk}^{(1)} : g = 1, \cdots, G; k = 1, \cdots, K\}$ *by maximizing* $f(\boldsymbol{N}_g|\alpha_{gk}, \mu_k^{(1)})$ *with respect to* α_{gk} *for each combination of gene* g *and cluster* k.

(ii) *E-step*: *Calculate the conditional expectation of* Z_{gk} *given data and parameters estimated from the* m*th step* $(\mu^{(m)}, \boldsymbol{p}^{(m)}, \alpha^{(m)})$, *where* $\mu^{(m)} = \{\mu_k^{(m)} : k = 1, \cdots, K\}$, $\boldsymbol{p}^{(m)} = \{p_k^{(m)} : k = 1, \cdots, K\}, \alpha^{(m)} = \{\alpha_{gk}^{(m)} : g = 1, \cdots, G;$ $k = 1, \cdots, K\}$. *To simplify the notation, we use* $\hat{Z}_{gk}^{(m)}$ *to denote the conditional expectation* $\mathrm{E}(Z_{gk}|\boldsymbol{N}, \mu^{(m)}, \boldsymbol{p}^{(m)}, \alpha^{(m)})$

$$\hat{Z}_{gk}^{(m)} = \frac{p_k^{(m)} f(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \mu_k^{(m)})}{\sum_l p_l^{(m)} f(\boldsymbol{N}_g|\alpha_{gl}^{(m)}, \mu_l^{(m)})}. \tag{4}$$

(iii) *M-step*: *Update the parameter estimates by*

$$\mu_k^{(m+1)} = \underset{\left\{\sum_i \mu_{ki}=0\right\}}{\operatorname{argmax}} \sum_g \hat{Z}_{gk}^{(m)} \log f\left(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \mu_k\right)$$

$$p_k^{(m+1)} = \sum_g \hat{Z}_{gk}^{(m)} \Big/ G$$

and

$$\alpha_{gk}^{(m+1)} = \underset{\alpha_{gk}}{\operatorname{argmax}}\, f\left(\boldsymbol{N}_g|\alpha_{gk}, \mu_k^{(m+1)}\right)$$

where $\hat{Z}_{gk}^{(m)}$ *is obtained from from step (ii).*

(iv) *Return to step (ii) or stop the iteration if change of the total log-likelihood is small.*

(v) *For each* $g = 1, \cdots, G$, *assign gene* g *to cluster* k *if* $k = \operatorname{argmax}_l \hat{Z}_{gl}$, *where* $\hat{Z}_{gl}$ *is obtained after the convergence of aforementioned steps.*

Note that Algorithm 1 not only assigns gene g to cluster k but also provides a measure of the uncertainty in the assignment by $1 - \hat{Z}_{gk}$. If clustering based on $\alpha_g + \beta_g$ is preferred, then we do not estimate α_{gk} but estimate α_k together with μ_k and corresponding calculations in step (i)–(iii) can be easily modified.

3.2 Initialization

It is well known that initialization of the cluster centers impacts both the speed of convergence and the outputs of the EM algorithm (Fraley and Raftery, 2002; Hall *et al*., 1999; Park *et al*., 2005). To tackle this problem, Arthur and Vassilvitskii (2007) proposed to pick the initial cluster centers from observations in a specific way such that they are well separated from each other with respect to some distance measure. Following this idea, rather than choosing K genes uniformly at random from all genes and using their expression profiles as the initial cluster centers, we only choose one cluster center uniformly at random and then set the additional centers gradually by selecting genes based on the distance between each gene and each of the selected centers. Here, the distance is measured by likelihood function.

Algorithm 2: Model-based Initialization for Cluster Centers.

(i) *Choose one gene randomly from all genes, and set the initial center for cluster 1,* $\mu_1^{(1)}$, *to be the maximum likelihood estimate (MLE) of* β_g *of the selected gene.*

(ii) *Given* m *center(s),* $\mu_1^{(1)}, \cdots, \mu_m^{(1)}$ *for* $1 \leq m < K$, *selected from previous steps, calculate the measure of the distance,* d_{gl}, *between each gene* g *and each previously*

selected cluster center $\mu_l^{(1)}$ *by*

$$d_{gl} = \log \frac{\max_{\alpha_g \in \mathcal{R},\ \sum \beta_{gl}=0} f(\boldsymbol{N}_g|\alpha_g, \beta_g)}{\max_{\alpha_g \in \mathcal{R}} f(\boldsymbol{N}_g|\alpha_g, \beta_g = \mu_l^{(1)})}$$

for $g = 1, \cdots, G; l = 1, \cdots, m$. *Then randomly select a gene with probability* $q_g = d_g^2 / \sum_{g'=1}^{G} d_{g'}^2$ *for* $d_g = \min\{d_{g1}, \cdots, d_{gm}\}$ *and set a new center* $\mu_{m+1}^{(1)}$ *as the MLE of* β_g *for the selected gene in this step.*

(iii) *Repeat step (ii) until K cluster centers are obtained.*

By the definitions of d_g and q_g in step (ii) of Algorithm 2, a gene is more likely to be selected if it is far away from all existing centers. Hence the K centers chosen by this algorithm are expected to be separated better than a set of centers that are randomly selected. Our simulation study shows that this algorithm improves the performance of EM algorithm (Section 4.4).

3.3 Other algorithms for model-based clustering

The EM algorithm does not guarantee global optimal solutions. Several stochastic algorithms have been proposed to reduce the risk of being trapped in local solutions. We describe two in this subsection and will examine their performances in our analysis. Both algorithms modify Equation (4) to calculate $\hat{Z}_{gk}^{(m)}$ in step (ii) of Algorithm 1.

(a) According to the deterministic annealing (DA) algorithm described in Rose (1998), the cluster in the m th iteration step is updated by

$$\hat{Z}_{gk}^{(m)} = \frac{p_k^{(m)} \left\{ f(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \mu_k^{(m)}) \right\}^{1/\tau_m}}{\sum_l p_l^{(m)} \left\{ f(\boldsymbol{N}_g|\alpha_{gl}^{(m)}, \mu_l^{(m)}) \right\}^{1/\tau_m}} \tag{5}$$

(b) The classification expectation maximization (CEM) algorithm with simulated annealing (SA) proposed by Celeux and Govaert (1992) updates the estimate of Z_{gk} by

$$\hat{Z}_{gk}^{(m)} = \frac{\left\{ p_k^{(m)} f(\boldsymbol{N}_g|\alpha_{gk}^{(m)}, \mu_k^{(m)}) \right\}^{1/\tau_m}}{\sum_l \left\{ p_l^{(m)} f(\boldsymbol{N}_g|\alpha_{gl}^{(m)}, \mu_l^{(m)}) \right\}^{1/\tau_m}} \tag{6}$$

Both algorithms use the annealing procedure with a sequence of preselected annealing rates ('temperatures', τ_m) decreasing to zero from a positive number. Apparently, when fixing $\tau_m = 1$, both algorithm updates the values of $\hat{Z}_{gk}^{(m)}$ the same way as the EM algorithm. Hence, Algorithm 1 can be viewed as a special case with a constant annealing rate $\tau_m \equiv 1$. As $\tau_m \to \infty$, we always get $\hat{Z}_{gk}^{(m)} = p_k$ for DA algorithm and $1/K$ for SA algorithm, which means that genes are assigned to each cluster totally randomly. On the other hand, as $\tau_m \to 0$ the randomness is gradually lost and we finally get $Z_{gk} = 0$ or 1, i.e. a hard cluster solution. Hence, τ_m determines the amount of randomness added in each step while searching for solutions. To apply these algorithms, we follow the suggestions of Rose (1998) and use $\tau_{m+1} = 0.9\tau_m$ with $\tau_1 = 2$.

For the SA algorithm proposed in Celeux and Govaert (1992), another difference from the EM algorithm (Algorithm 1) is that, before updating parameter values in the M-step, each gene is assigned to a cluster based on one random draw from a multinomial distribution with probabilities $\hat{Z}_{gk}^{(m)}$ as calculated by Equation (6).

3.4 Model-Based Hybrid-Hierarchical Clustering Algorithm

So far, we have assumed that the number of clusters, K, is predetermined. For a real data analysis, this quantity often needs to be estimated. There are different methods that can be applied to estimating K. For instance, choose the K that minimizes the Akaike information criterion (AIC) for the mixture model. Alternatively, instead of choosing a single value of K for the clustering analysis, we can build a hierarchical tree of clusters. The hierarchical structure of the clusters provides information about the relationships of clusters and allows flexibility of obtaining different number of clusters by cutting the tree at different levels.

There can be tens of thousands of genes from RNA-seq data, and treating each gene as the smallest cluster at the bottom of the tree requires intensive computation. To speed up the calculation, we propose to use agglomerative (bottom-up) strategy starting with K_0 clusters, where K_0 is a number relatively large to allow enough resolution but far less than the number of genes, G. The initial K_0 clusters can be obtained by the model-based clustering algorithms described in the previous subsections. In each of the following steps, two clusters are merged if the 'distance' between them is the smallest among all possible pairs. Finally after $K_0 - 1$ steps, all genes belong to a single cluster and the hierarchical tree is built up. Such an algorithm has been called hybrid-hierarchical (HH) clustering algorithm (Vaithyanathan and Dom, 2000; Zhong and Ghosh, 2003). Here, the term 'hybrid' is used to point out that the HH algorithm combines the starting steps that obtain K_0 clusters using non-hierarchical methods and the merging steps that are similar to ordinary hierarchical clustering.

After the mth $(0 \le m < K_0)$ merging step, we denote the $K_0 - m$ clusters by disjoint sets $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{K_0-m}$, and calculate the distance between two clusters, say $\mathcal{G}_k$ and $\mathcal{G}_l$, by Equation:

$$D(\mathcal{G}_k, \mathcal{G}_l) = \log \frac{\prod_{g \in \mathcal{G}_k} f\left(\boldsymbol{N}_g|\alpha_g^{(k)}, \mu_k\right) \prod_{g \in \mathcal{G}_l} f\left(\boldsymbol{N}_g|\alpha_g^{(l)}, \mu_l\right)}{\prod_{g \in \mathcal{G}_k \cup \mathcal{G}_l} f\left(\boldsymbol{N}_g|\alpha_g^{(kl)}, \mu_{(kl)}\right)} \tag{7}$$

where $\alpha_g^{(k)}$ and μ_k maximize the likelihood $f(\boldsymbol{N}_g|\alpha_g, \mu_k)$, and $\mu_{(kl)}$ is the center of the cluster formed by merging $\mathcal{G}_k$ and $\mathcal{G}_l$. This distance is the reduction of total log-likelihood from before to after the mergence. Obviously, merging clusters with the minimal distance defined in (7) aims to achieve the maximum log-likelihood in each step (Fraley, 1999; Meila and Heckerman, 2001).

4 SIMULATION STUDY

We conducted simulation studies to compare model-based clustering methods with other methods, including K-means and SOM, which have been popularly used in microarray data analysis and could also be applied to analyzing RNA-seq data.

We first describe how data were generated in Section 4.1 and present the criteria used to evaluate the clustering performance in Section 4.2. Then we check the validity of treating the estimated dispersion parameter ϕ_g as known for NB models in Section 4.3 and evaluate the model-based initialization algorithm (Algorithm 2) versus random initialization in Section 4.4. Finally, in Section 4.5, we compare our proposed algorithms with others.

4.1 Data simulation

We considered an experiment with three treatment groups and three replicates for each treatment group. This is a case easily encountered in real data analysis. Suppose that there were $K = 7$ different expression patterns across three treatments and the cluster centers were characterized by $\mu_k = \eta_\mu \delta_k$, where η_μ determined the magnitude of gene expression changes across treatments and $\delta_k = (\delta_{k1}, \delta_{k2}, \delta_{k3})$ described the pattern of changes for cluster k, for $k = 1, \cdots, K$. A larger η_μ means larger distances between the centers and better separation of clusters. The distinct profiles characterized by $(\delta_{k1}, \delta_{k2}, \delta_{k3})$ are listed as follows:

Cluster k	1	2	3	4	5	6	7
δ_{k1}	−1	−1	0	0	1	1	0
δ_{k2}	0	1	−1	1	−1	0	0
δ_{k3}	1	0	1	−1	0	−1	0

For the first cluster, the expression of genes increases from the first treatment group to the second one and increases further for the third treatment group. For the second cluster, the expression increases from first treatment group to the second one but then decreases for the third group. Note that the last cluster has a mean profile identically zero and this cluster corresponds to the group of genes that are non-differentially expressed across treatments. Although only identified differentially expressed genes are typically included in the cluster analysis, there could be false positives on the list of identified genes. For the simulation study, we included this cluster of non-differentially expressed genes to make our simulation more general and did not expect this to affect the relative ranking of the evaluated methods.

RNA-seq data for $G = 10\,000$ genes were simulated for each dataset according to the following regime. For each $g = 1, \cdots, G$, $\boldsymbol{Z}_g^0 = \{Z_{gk}^0 : k = 1, \cdots, 7\}$ was drawn independently from a multinomial distribution with equal probabilities, where $Z_{gk}^0 = 1$ means gene g belongs to cluster k and $Z_{gk}^0 = 0$ otherwise. Given $Z_{gk}^0 = 1$, the gene expression profile was simulated according to $\beta_g = \mu_k + \epsilon_g$, where $\mu_k = \eta_\mu \delta_k$ as described earlier in the text and $\epsilon_g = (\epsilon_{g1}, \epsilon_{g2}, \epsilon_{g3})$ added fluctuation around cluster center μ_k specifically for gene g. We sampled ϵ_{gi} for $i = 1, 2, 3$ from $\eta_\mu \eta_\epsilon N(0, 0.2^2)$, where η_ϵ controlled the level of fluctuation relative to the cluster center $\eta_\mu \delta_k$. The overall mean expression level α_g was drawn from $\eta_\alpha N(4, 1)$, where η_α controlled the magnitude of average expression level. The dispersion parameter ϕ_g was simulated from η_ϕGamma$(0.75, 2)$, where Gamma$(0.75, 2)$ is a gamma distribution with mean $0.75/2$ and variance $0.75/2^2$. Changing the value of η_ϕ allowed different levels of dispersion. Specially, $\eta_\phi = 0$ corresponds to the Poisson model, which is the limiting case of NB model as the dispersion approaches zero. The normalization factor s_{gij} was generated from $N(0, 1)$. Given these parameters, the gene expression count N_{gij} was generated from the NB model with expectation $\exp(s_{gij} + \alpha_g + \beta_{gi})$ and dispersion ϕ_g.

Once the dataset was simulated, we treated all parameters except s_{gij} as unknown to resemble a real experiment. The values of $\eta_\mu, \eta_\epsilon, \eta_\alpha$ and η_ϕ were varied to create different simulation settings, and 100 datasets were independently simulated for each setting.

To test the robustness of our model, we also simulated data according to a generalized linear mixed model (GLMM) $N_{gij} \sim NB\big(\exp(s_{gij} + \alpha_g + \mu_{ki} + \epsilon_{gi} + \gamma_{gij}), \phi_g\big)$. Here, we added a random effect γ_{gij} to the expected expression, where γ_{gij} is specific for each combination of gene and sample. γ_{gij} was drawn from a normal distribution $\eta_\mu \eta_\epsilon N(0, 0.1^2)$. With this GLMM model, we have overdispersed data compared with the NB model that we assume in (3). The results based on data simulated from both models [GLMM and the NB model with expectation $\exp(s_{gij} + \alpha_g + \beta_{gi})$] are similar, and our conclusions are the same. So we only present the results based on the NB model.

4.2 Assessment of performance

We assessed the performances of different clustering approaches by comparing the resulting partitions with the original partition of genes defined by $\boldsymbol{Z}^0 = \{\boldsymbol{Z}_g^0 : g = 1, \cdots, 10000\}$. A better performance is indicated by more agreement between the two partitions. The following three statistics were used to evaluate the agreement. For all the three statistics, higher values indicate better performance.

(1) *Pairwise sensitivity*: the proportion of pairs of genes (objects) that are clustered together among all pairs that had the same original assignment (Booth *et al.*, 2008; Woodard and Goldszmidt, 2011).

(2) *Pairwise specificity*: the proportion of pairs of genes (objects) that are clustered to different groups among all pairs that had different original assignment (Booth *et al.*, 2008; Woodard and Goldszmidt, 2011).

(3) *Normalized mutual information (NMI)*: MI is used in information theory to measure the amount of information one random variable contains about another, or equivalently, the reduction in the uncertainty of one due to the knowledge of the other. Here, MI is used to quantify the shared information between the true partition and the clustering result. See Strehl and Ghosh (2002) for the explicit formula for calculation using the contingency table formed by the two partitions. MI value is high if there is strong dependence (more shared information) between the two partitions, and is close to zero otherwise. Because there is no upper bound for MI, its normalized version ranging from 0 to 1 is often desirable for easier comparison (Strehl and Ghosh, 2002).

4.3 Validation of estimating dispersion parameters

We estimated the dispersion parameters ϕ_g and treated them as if they were true values when applying the model-based clustering algorithms. However, it is challenging to obtain good estimates of dispersion parameters due to the small number of replicates in

RNA-seq data. To examine the impact of the estimated parameters on cluster analysis, we compared the model-based clustering methods using estimated values for ϕ_g versus that using the input (true) values used to simulate the counts.

Figure 1a plots the values of sensitivity, specificity and NMI for different clustering approaches over a range of η_ϵ values used to simulate RNA-seq data, whereas other parameters η_μ, η_α and η_ϕ were fixed at 1. As shown in Figure 1a, when $K=7$ and at the same level of η_ϵ, the MB-EM algorithms using true and estimated dispersions perform indistinguishably as shown in Figure 1a. In practice, the true number of clusters is unknown, and we might apply a different number in cluster analysis, say $K=10$. Still, the clustering results from using true and estimated dispersions are almost the same. We also varied parameters η_α, η_μ and η_ϕ one at a time while keeping others fixed at 1 to generate RNA-seq datasets. The difference between using true and estimated dispersions were small at most of the parameter settings (see Supplementary Fig. S1). Consequently, all results presented later were obtained using estimated dispersion parameters just like how we analyze real data.

It is worth pointing out that we *cannot* conclude that the results for $K=10$ are better than that for $K=7$, though the specificity scores for the former are higher. Comparing the sensitivity or specificity scores is not meaningful when the numbers of clusters are different. For an extreme example, the sensitivity will always be 1 when $K=1$ because all gene pairs that had the same original assignment will be clustered together. Similarly, when choosing K as high as 10 000, the specificities will always be 1.

4.4 Comparison of initialization algorithms

In Figure 1b, we compared the initialization effects on the MB-EM clustering results. Our proposed model-based algorithm (Algorithm 2) and random initialization were examined. Though initialization using true cluster centers is not applicable in practice, we also included it in the comparison as a gold standard to evaluate the other two initialization methods. Figure 1b clearly illustrates that the model-based initialization performs much better than random initialization by giving higher evaluation statistics for all parameter settings in simulation. In many cases, the model-based approach generated results similar to those when the true cluster centers were applied for initialization. Results for other simulation settings are presented in Supplementary Figure S2.

4.5 Comparison of MB cluster algorithms with others

We proposed EM algorithm (Algorithm 1) to perform model-based clustering. However, it is possible that the resulting partition from EM algorithm is not a global optimum. Hence, two stochastic versions, DA and SA algorithms, are described in section 3.3 to reduce such risk. In this section, we compare these slightly differing algorithms, whereas all three were initialized with the same set of cluster centers chosen by Algorithm 2. First, we did cluster analysis with the true number of clusters, $K=7$. Figure 1c and Supplementary Figure S3 suggest that all three algorithms perform almost the same. We also analyzed the same datasets with $K=10$ (Fig. 1c and Supplementary Fig. S4). Interestingly, Supplementary Figure S4 shows that the SA algorithm typically achieves the highest sensitivity, whereas the DA algorithm gains in terms of specificity. If practitioners are more interested in sensitivity, getting groups of genes with similar profiles, then the SA algorithm is recommended. If separating genes with different profiles is more of interest, then DA algorithm can be applied.

We also compared the proposed algorithms with K-means and SOM, two methods that have been popularly applied to microarray analysis and can potentially be applied for RNA-seq data. To cluster gene expression profiles, K-means and SOM were applied to cluster the MLEs obtained based on the NB model, i.e. the mean profile of normalized RNA-seq data across replicates for each gene. Plots in Figure 1c and d and Supplementary Figures S3 and S4 show that, evaluated by all three criteria, the model-based algorithms perform obviously better than K-means and even better than SOM. Note that our simulation settings include Poisson model, which is a special case when the dispersion parameter is set to be zero. We also did more simulations with Poisson model and the results are similar to what are shown here.

4.6 Choosing the number of clusters

One important question in the implementation of model-based cluster analysis for real data is to choose the number of clusters, K. Here, we evaluated the AIC. For given K, we can calculate the likelihood L by (3) and the AIC by $-2(\log L - n_p)$, where $n_p = G(K+1) + KI - 1$ is the number of parameters in the model. A low value of AIC indicates a better clustering result. As shown in Figure 2a, the AIC identified the true number of clusters being optimal.

5 REAL DATA ANALYSIS

Li *et al.* (2010) studied the maize leaf transcriptome using Illumina Genome Analyzer 2. The dataset quantifies transcript abundance of four sections along a leaf developmental gradient, with two biological replicates for each section. Using generalized linear model analysis based on NB distribution, we found that 12 631 genes were differentially expressed across the four sections. Li *et al.* (2010) normalized the count data by calculating the values of reads per kilobase of exon model per million mapped reads (RPKM), a popular quantification method proposed by Mortazavi *et al.* (2008). In this section, on log-transform and mean-center the RPKM values for each gene, we obtained the log fold change estimates of the expressions relative to the average expression of each gene. To these log fold change estimates, we applied both the K-means, which has been used in Li *et al.* (2010), and the SOM algorithms. We also present results from the model-based clustering algorithms for the untransformed count data based on NB model. One advantage of the model-based approaches is that the Poisson or NB model can handle genes with low counts easily. When sequencing depth is low, there may be many genes with low counts or zero counts in some replicates or treatment groups. However, this will induce problems in the log-transformation, which is typically done before applying K-means method. The following numerical results also show that our proposed method provides better clusters than both K-means and SOM algorithms.

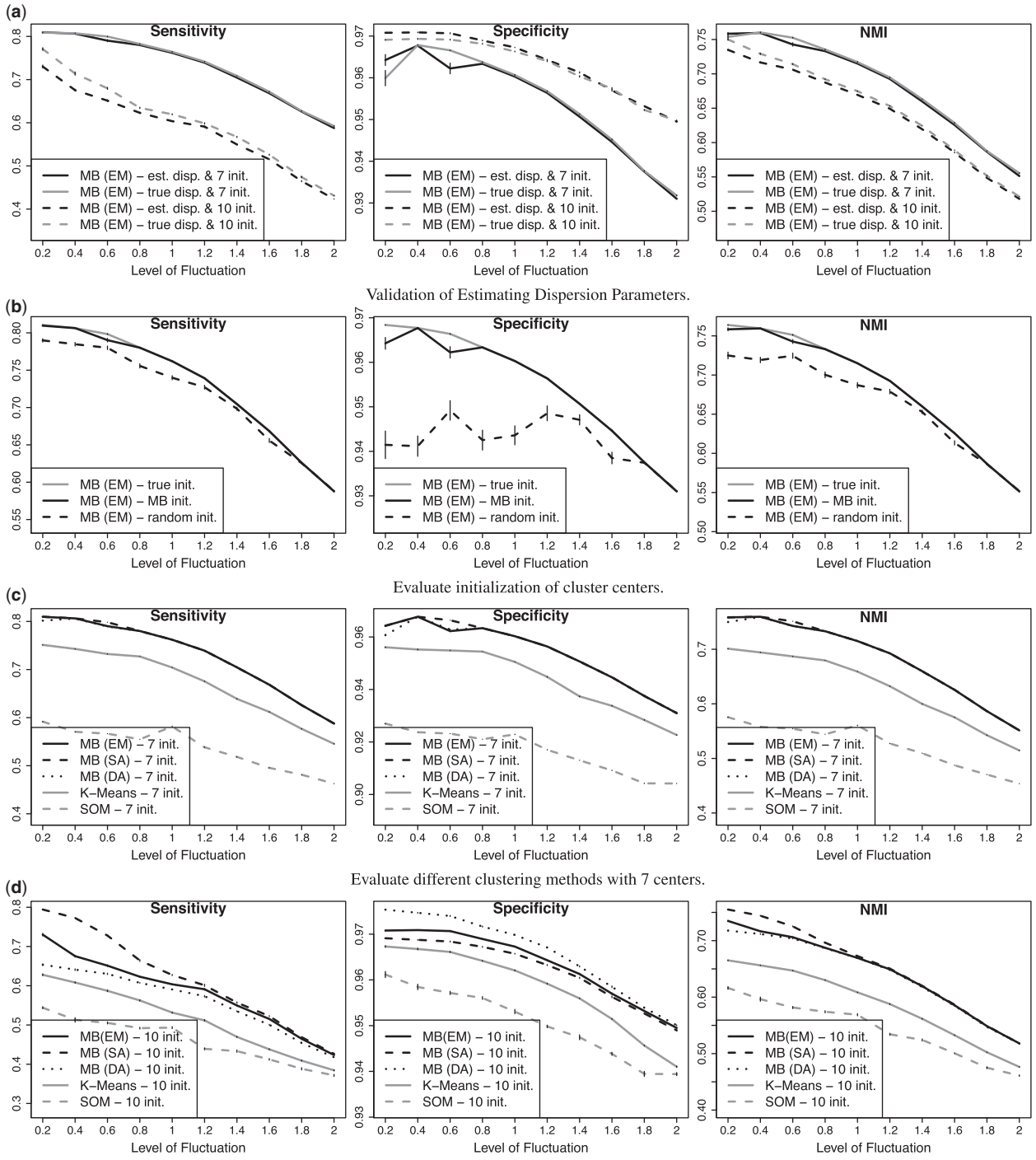

Fig. 1. Simulation results. The level of fluctuation, η_ϵ, was increased from 0.2 to 2. See Supplementary Figures S1–S4 for more results when adjusting the level of dispersion η_ϕ, the magnitude of log-FC η_μ and the overall expression η_α. For each parameter setting, the clustering results from 100 datasets, each containing 10 000 genes simulated, were evaluated by sensitivity, specificity and NMI, and the scores were averaged across the 100 datasets. The length of each vertical bar on the lines represents the standard error. Note that some standard error bars are too small to be seen from this graph. (**a**) The MB-EM algorithms using true (true disp.) and estimated dispersion (est. disp.) parameters were compared for either 7 or 10 initialization centers (7 or 10 init.). (**b**) The initialization with model-based algorithm (Algorithm 2, MB init.) and initialization with randomly picked objects (random init.) are compared with initialization with true cluster centers (true init.). (**c** and **d**) Comparison of the three model-based methods (EM, DA and SA algorithms) with the non-MB methods includes K-means and SOM. All are initialized by the same 7 or 10 cluster centers chosen by Algorithm 2

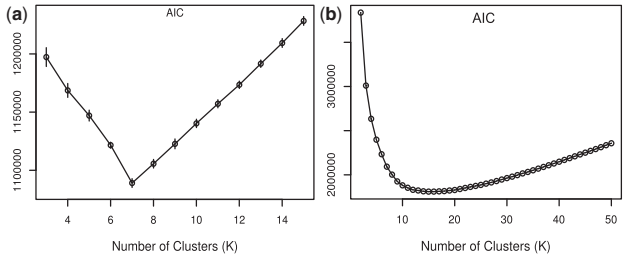

Fig. 2. Number of clusters. (**a**) The clustering results in the simulation study were evaluated by the AIC. Under the simulation setting $\eta_\alpha = \eta_\mu = \eta_\epsilon = \eta_\phi = 1$, 100 independent datasets were simulated. Results are averaged over the 100 datasets, and the length of the vertical bar at each point is the standard error of the mean of the score. (**b**) The clustering results for the maize data were evaluated by the AIC (see Section 5 for real data analysis)

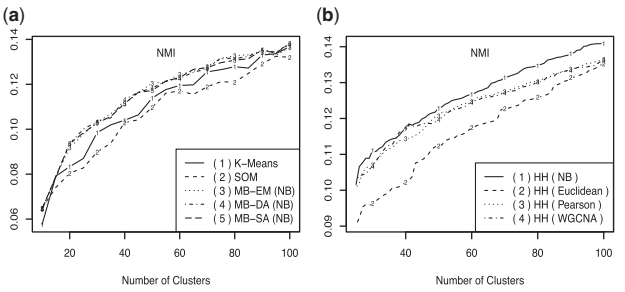

Fig. 3. Clustering results for the maize dataset. (**a**) We compared our proposed model-based algorithms (EM, DA and SA) with the K-means and SOM methods. (**b**) MB-HH is compared with hierarchical clustering based on Euclidean distance, Pearson correlation and similarity function in WGCNA. They all start from 100 clusters obtained using their corresponding distance measures

As we expect that the genes within the same functional category have correlated expression patterns and thus more likely to be grouped together, a clustering result can be evaluated by checking its concordance with the functional categories. Gene annotations were obtained from Mapman as described in Li *et al*. (2010). Excluding categories that contain <5 or >500 genes, we ended with 306 non-overlapping categories with a total of 5002 genes. Because these annotations are independent to the clustering processes, the evaluation is not biased toward any clustering method and data model.

We first used $K = 100$ to cluster genes using both our model-based method and the K-means method. The reason that we chose $K = 100$ is because we presume that more clusters can give better resolution of expression trends to the grouped genes with the 306 Mapman categories. We are interested in genes that show monotonic expression profiles along the leaf gradient, and we found that genes in clusters 14, 18 and 21, which are the three biggest clusters resulting from our model-based method, show a monotonic decreasing pattern from base to tip, which may help us to discover the biology that distinguishes base from other sections (Supplementary Table in excel file). We found that 23 genes in cell wall functional category according to Mapman annotation are grouped into cluster 21. However, these genes are scattered around different clusters obtained from the K-means method. The cell wall functional category totally includes 165 genes. We noticed that in model-based method, the cell wall related genes are enriched in cluster 14 (15 genes in cluster 14) and 18 (15 genes in cluster 18), in addition to cluster 21 (23 genes in cluster 21), which all represent the higher gene expression in base. However, these genes were scattered into 23 clusters obtained from K-means method, and there is no cluster identified by K-means that includes >10 genes from this gene category. Only by looking at these three clusters from model-based method, we can clearly conclude that there was an active cell wall metabolism at the basal part of developing leaf, which is not easy to detect using the K-means method. In addition, cell organization and DNA synthesis/chromatin structure pathways were also enriched in cluster 21 in model-based method, which suggested active cell construction and DNA replication in the leaf base, and this is consistent with the active cell wall metabolism in the basal part of leaf. All these biological events were easily identified by the model-based method, but not the K-means method.

To obtain a more quantitative analysis, we measured the concordance between clustering results and gene functional categories by NMI. We performed cluster analysis with $K = 10, 15, 20, \cdots, 100$ clusters for all five methods, including SOM, K-means and the three model-based algorithms. Figure 3a shows that the model-based algorithms outperform SOM and K-means for all K values in terms of NMI. We then applied the HH clustering as described in Section 3.4, starting from $K_0 = 100$ clusters obtained using the corresponding distance measures. We also applied hierarchical clustering using average linkage based on Euclidean distance, Pearson correlation and the adjacency (similarity) function in *weighted gene co-expression network analysis* (WGCNA) proposed by Zhang and Horvath (2005). Our proposed HH method generated higher NMI scores (Fig. 3b) than the other three hierarchical methods. Examples of the clustering results for $K = 20$ and hierarchical structures for the model-based hybrid-hierarchical clustering algorithm (MB-HH) clusters are plotted in Figure 4 and Supplementary Figures S5 and S6. These plots show that the EM algorithms result in much cleaner expression patterns than the clusters obtained from either K-means or SOM algorithm.

We also used the AIC criterion based on NB models, similarly as in Section 4.6, to decide the number of clusters. We found $K = 15$ is the optimal number of clusters by AIC (Fig. 2b).

6 DISCUSSION

In this article, we derived clustering algorithms based on finite mixture of Poisson or NB models. We proposed an EM algorithm with model-based initialization, and show this initialization method greatly improves the performance of the EM clustering. Compared with heuristic algorithms such as K-means method, our method has the following advantages: First, we build our approach of clustering RNA-seq data based on more appropriate probabilistic models such as Poisson and NB distributions. Owing to the nature of RNA-seq technology, the observed count data are discrete and skewed. Poisson model has been shown to fit well to data without biological replicates (Marioni *et al*., 2008) and NB model to data with biological replicates (Anders and

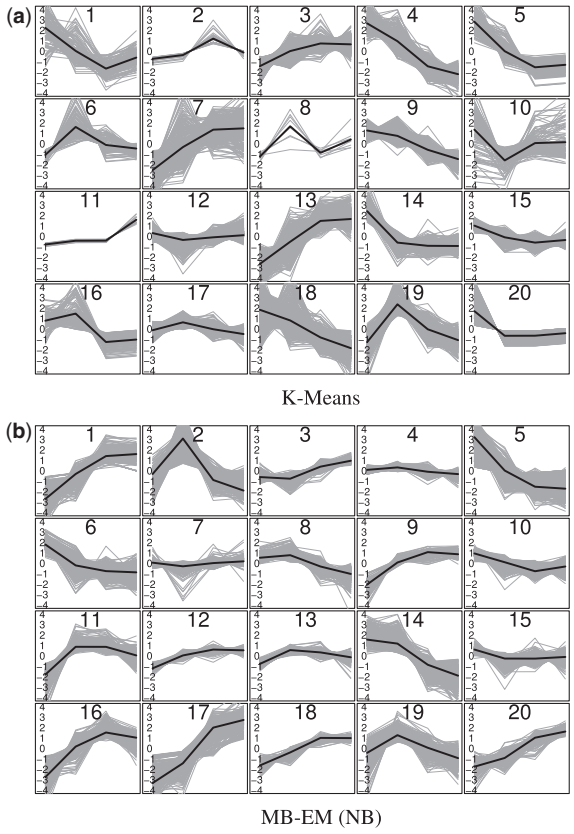

Fig. 4. Real data analysis: (**a**) the result from K-means algorithm using Euclidean distance; (**b**) the result from EM algorithm based on NB model. The gray lines correspond to the gene expression patterns estimated by method of moments, and the black lines plot the cluster centers

Huber, 2010). Second, we demonstrated through both simulation studies and real data analysis that our proposed algorithms outperformed heuristic methods such as K-means and SOM, which have been popularly applied to cluster gene expressions from microarray and can also be applied to RNA-seq data. Third, we propose the MB-HH that allows flexibility in applying our method. Finally, our method provides a unified way to select the number of clusters. Using our models, we can evaluate the model selection criterion, AIC, and decide the number of clusters to use. Although our method is illustrated with analysis of data from completely randomized design, other more complex designs can be handled by appropriately modifying our model (1) and likelihood (3).

Funding: This research was supported in part by the National Science Foundation (NSF) Grants (No. IOS-0701736 and IOS-1127017).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Arthur,D. and Vassilvitskii,S. (2007) K-means++: the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. pp. 1027–1035.

Booth,J. *et al.* (2008) Clustering using objective functions and stochastic search. *J. R. Stat. Soc. Series B*, **70**, 119–139.

Bullard,J. *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**, 94.

Celeux,G. and Govaert,G. (1992) Ea classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, **14**, 315–332.

Fraley,C. (1999) Algorithms for model-based gaussian hierarchical clustering. *SIAM J. Sci. Comput.*, **20**, 270–281.

Fraley,C. and Raftery,A. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Hall,L. *et al.* (1999) Clustering with a genetically optimized approach. *IEEE Trans. Evol. Comput.*, **3**, 103–112.

Li,P. *et al.* (2010) The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.*, **42**, 1060–1067.

Marguerat,S. *et al.* (2008) Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.*, **36**, 1091–1096.

Marioni,J.C. *et al.* (2008) Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

McLachlan,G. (1997) On the em algorithm for overdispersed count data. *Stat. Methods Med. Res.*, **6**, 76–98.

Meila,M. and Heckerman,D. (2001) An experimental comparison of model-based clustering methods. *Mach. Learn.*, **42**, 9–29.

Metzker,M. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by Rna-seq. *Nat. Methods*, **5**, 621–628.

Park,H. *et al.* (2005) Evolutionary fuzzy clustering algorithm with knowledge-based evaluation and applications for gene expression profiling. *J. Comput. Theor. Nanosci.*, **2**, 1–10.

Ressom,H. *et al.* (2003) Clustering gene expression data using adaptive double self-organizing map. *Physiol. Genomics*, **14**, 35–46.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of Rna-seq data. *Genome Biol.*, **11**, R25.

Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, **9**, 321–332.

Robinson,M.D. *et al.* (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rose,K. (1998) Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, **86**, 2210–2239.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles - a knowledge reuse framework for combining partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Vaithyanathan,S. and Dom,B. (2000) Model-based hierarchical clustering. In: *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. pp. 599–608.

Wang,L. *et al.* (2010) Exploring plant transcriptomes using ultra high-throughput sequencing. *Brief. Funct. Genomics*, **9**, 118–128.

Wang,Z. *et al.* (2009) Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 53–67.

Witten,D.M. (2011) Classification and clustering of sequencing data using a poisson model. *Ann. Appl. Stat.*, **5**, 2493–2518.

Woodard,D. and Goldszmidt,M. (2011) Model-based clustering for online crisis identification in distributed computing. *J. Am. Stat. Assoc.*, **106**, 49–60.

Yeung,K. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Zhang,B. and Horvath,S. (2005) General framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**. Article 1.

Zhong,S. and Ghosh,J. (2003) A unified framework for model-based clustering. *J. Mach. Learn. Res.*, **4**, 1001–1037.