Revision for "Model-Based Clustering for RNA-Seq Data" by Yaqing Si, Peng Liu, Pinghua Li and Thomas P. Brutnell

The authors wish to thank Associate Editor Michael Brudno and the three anonymous referees for their careful review of our paper. We have modified the manuscript to address all comments and suggestions, and we believe this process has resulted in a further improved paper. A point-by-point response to all the comments raised during the review process is provided below. Reviewer comments are italicized. Our response in standard text follows each comment.

Response to Reviewer 1

The authors of this paper derived model-based clustering algorithms based on appropriate probability models for RNA-seq data and compared the performance of these model-based approaches with heuristic algorithms including the K-means and self-organizing map (SOM) methods. The manuscript is well written in general. The methodologies are clearly described, followed by simulation study and real data analysis.

Response: Thank you for reviewing our paper, and we are glad that you think this manuscript is well-written in general.

Major Comments:

1. *When the authors compared their proposed algorithms with K-means and SOM methods, data were simulated under models that satisfy assumptions of their model-based clustering algorithms. Is it fair for comparisons? How to check if real data follow distributions that authors assume, e.g. Poisson or Negative Binomial or in the presence of overdispersion? If data were simulated from other models or mixed models, how about the performance of these methods?*

   Response: The model used to simulate data is the same as the model we used to derive the model-based clustering algorithm except that we added the gene-specific fluctuation around the center of the clusters when we simulated data. Nevertheless, the two models are similar. The Poisson distribution has been shown to fit RNA-seq data well by goodness-of-fit tests when there is technical replication (Marioni et al., 2008, Bullard et al., 2010). When there is biological replication, more variation is expected, and the Negative Binomial (NB) distribution is a reasonable model for RNA-seq data (Anders and Huber, 2010, Hardcastle and Kelly, 2010, etc.). Popular R packages such as `edgeR, baySeq, DESeq` base their methods on NB distributions. Hence, we choose such a common strategy to simulate data.

   We agree with the referee that it would be better to check whether real data follow the assumed distributions. For the case of RNA-seq data analysis, this is an interesting and challenging question because we have tens of thousands of genes but only very small number of replicates, i.e., we have a so-called "small n, large p" case. We are aware that some research is going on to address this question but no published paper can be found yet. More discussions about this question is beyond the scope of our paper. Instead of proposing a method for diagnostic tests, we evaluated the robustness of our method under a different model. More specifically, we simulated counts using a generalized linear mixed model (GLMM) as follows.

$$N_{gij} \sim NB\left(\exp(s_{gij} + \alpha_g + \mu_{ki} + \epsilon_{gi} + \gamma_{gij}), \phi_g\right).$$

   We use the index g for genes, i for treatment, j for replicate, and k for cluster. Compared with the model $N_{gij} \sim NB\left(\exp(s_{gij} + \alpha_g + \mu_{ki} + \epsilon_{gi}), \phi_g\right)$ in our manuscript,

we added a random effect γ_{gij} to the expected expression, where γ_{gij} is specific for each combination of gene and sample. We drew γ_{gij} from a Normal distribution $\eta_\mu\eta_\epsilon N(0, 0.1^2)$ (for a comparison, ϵ_{gi} is generated from $\eta_\mu\eta_\epsilon N(0, 0.2^2)$). With the above model, we have over-dispersed data compared with the NB model that we assume in the manuscript. We found that the results (Figures 1-4 at the end of this response letter) are very similar to what we present in our manuscript, and our conclusion stays the same: our proposed methods still show obvious advantages over heuristic clustering algorithms such as K-means and SOM. We now describes this result in the paragraph above section 4.2 in our revised manuscript.

As we have pointed out in the manuscript, our algorithms are not limited to Poisson or NB distribution. The notation $f(\boldsymbol{N}|\alpha, \boldsymbol{\beta})$ is very generic so that α is the average gene expression and the vector $\boldsymbol{\beta}$ represents the pattern of expression changes. In case another distribution is affirmed to be more appropriate for an RNA-seq dataset, the model-based clustering algorithm can be extended to the new distribution by using the corresponding likelihood function for $f(\boldsymbol{N}|\alpha, \boldsymbol{\beta})$.

2. *As these methods require a priori the number of clusters (K), the authors compared the performance of these methods with different number of clusters and also provided criteria of choosing the number of clusters including AIC and SD. It was nice to see that AIC an SD consistently output the true number of clusters in the simulation study. In the real data analysis, however, AIC suggested K=15 while SD suggested K=2. It is lack of further discussion about what K is appropriate in such situation.*

Response: Thank you for pointing this out. We recommend to use the AIC criterion to choose the number of clusters (K) for two reasons. First, AIC is calculated using the log-likelihood of the mixture model, and hence is directly connected to the model-based algorithms. Second, To cluster thousands of genes, $K = 2$ results in large clusters with mixed patterns. It provided worse separation of interesting patterns compared with $K = 15$. To avoid confusion to practitioners and to save some space for the manuscript (because we added more literature review and discussion about the real data analysis), we now only use the AIC criterion and have an updated version of Figure 2 in the revised manuscript.

3. *It was surprised to see from Fig. 1(c) that three algorithms performed exactly the same with higher level of fluctuation. What might be the rationale behind it?*

Response:

With the model-based initialization, the EM, DA and SA algorithms perform indistinguishable on Figure 1(c) when the level of fluctuation is high, although they are not exactly the same because some genes, though very few, were assigned differently by the three algorithm.

We think a better mixing of genes is the reason why the three algorithms are more similar when the fluctuation level is high. Imagine the extreme case when the gene fluctuation η_ϵ is so large that $\boldsymbol{\epsilon}_g$ dominates the $\boldsymbol{\mu}_k$ so that all genes are well mixed. In such a case, we expect that it is less likely to be trapped in local optimum, and thus the two stochastic versions of EM algorithm (DA and SA) perform more similarly to the EM algorithm when using the same set of well selected initial points.

Minor Comments:

1. *On Page 4 line 27-28: It is not clear what does "one observation" mean when authors say "each gene is assigned to a cluster based on one observation simulated from a .".*

   Response: Thanks for letting us know it. We meant that one random draw from a multinomial distribution. We have changed that sentence to "each gene is assigned to a cluster based on one random draw from a multinomial distribution...".

2. *Legend in figure 1 needs to be explained. For example, what does "-est.disp. and 7 init." mean? Readers could guess but would appreciate if it can be addressed clearly.*

   Response: Thanks for pointing this out. We have added more explanation to the legend in Figure 1, and explained the abbreviations.

3. *It looks like some methods have standard deviations (vertical bars) plotted (in fig1) but some do not. Was it because those sds are too small hence "invisible" in the plot? Again, adding this information will reduce some confusion for readers.*

   Response: Yes, you are correct that some standard errors are too small to be visualized on the plots. We have added a remark to the figure legends. "Note that some standard error bars are too small to be seen from this graph."

4. *On page 7 line 13: "(MLE) of in the NB model", any typo or missing word between "of " and "in"?*

   Response: Thanks for pointing this out. We have changed this part to "(MLE) obtained based on the NB model".

Response to Reviewer 2

I found the paper to be both well-written and well-developed, with a thorough discussion of the overall method, although, as I'll detail below, I would like to see an expanded discussion of the results from the analysis of the real data.

I should clarify that I do believe that the approach that the authors present is novel, and is certainly novel to be used with RNA-seq data.

Response: Thank you for taking time to review our paper. We are glad that you think this paper is well-written and well-developed, and that you believe that this approach is novel. And we appreciate you bring out valuable comments about this work.

Major Comments:

1. *... the authors really should discuss this (the paper by McLachlan (1997)) review in their introduction, and explain how their method distinguishes itself from the ideas that it presents.*

   Response: Thank you for providing the reference McLachlan (1997). Both **?** and our paper employ EM algorithm to fit mixture models for count data. However, there are several major difference from our manuscript and this reference paper.

   (a) The research goals are very different between the two. **?** discusses the EM algorithm for finite mixture model so that a better distribution is used to handle overdispersed count data. Our purpose is cluster analysis.

   (b) **?** models univariate responses. We aim to cluster gene expression profile (a vector observation) into different groups.

   (c) **?** takes a GLM or GLMM setting with poisson or binomial distribution. We start with Negative Binomial distribution, which is more commonly used to model RNA-seq data.

   Thank you for your suggestion of reviewing this paper. We added this reference when we introduce EM algorithm in Section 3.1. It is brief due to space limitation. We quote the corresponding part here for your convenience to check it: "**?** describes an EM algorithm to fit overdispersed univariate count data in Poisson regression and logistic regression setting. Here, we derive an EM algorithm (ALG. 1) for clustering RNA-seq gene expression profile with a mixture of Poisson or NB models. "

   Following your suggestion to expand the literature review, we searched for clustering methods for count data or next-generation sequencing data. The most closely related paper to ours is Witten (The Annals of Applied Statistics, 2011). This paper discusses classification and clustering of samples (experimental units) using RNA-seq

data based on Poisson model, which is for a different goal from ours. We aim at clustering genes based on their expression profiles across treatments instead of clustering samples based on all genes' expression within each sample. In addition to the different goals, the clustering methods of Witten's and ours are different. Witten (2011) applies a hierarchical clustering algorithm based on dissimilarity measure based on Poisson likelihood ratio statistic, and we use model-based clustering methods based on finite mixture models of Poisson or Negative Binomial models. Nevertheless, we added this reference into the introduction on page 2 to offer readers a more comprehensive review of the current literature on clustering methods for RNA-seq data.

2. *In addition, it would behoove the authors to provide a more thorough description in the introduction of the clusters that their method seeks to identify.*

   Response: Thank you for your suggestion. We have added more description in the introduction section about what kinds of clusters that we aim to identify. We quote the addition in the first paragraph on page 2 here: "In this paper, 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."

3. *While it makes intuitive sense that a cluster of genes (i.e. the cluster centroid) should also follow a NB distribution, it is not clear to me that the distribution of genes about that centroid should also be distributed via a NB distribution.*

   Response: Conditional on knowing the cluster ID, we assume that genes follow NB distributions while genes in the same cluster share the same expression pattern ($\boldsymbol{\beta}_g = \boldsymbol{\mu}_k$). This conditional likelihood $f(\boldsymbol{X}_g|\alpha_g, \boldsymbol{\beta}_g = \boldsymbol{\mu}_k)$ is what we use in the EM algorithm. Marginally, genes follow a finite mixture model which allows extra variation and more flexibility than a single NB model. Such model used for cluster analysis has been employed in multivariate normal case where objects in the same cluster share the same mean. A subtle difference is that that our model allow different overall mean expression (α_g's) for genes belong to the same cluster. We focus on the differential expression pattern $\boldsymbol{\beta}_g$ because our biological collaborators are often interested in this. As mentioned in the paper, we could have $\alpha_g = \alpha_k$ and $\boldsymbol{\beta}_g = \boldsymbol{\mu}_k$ so genes in the same cluster have the same mean as in the normal case.

4. *As best I can tell, based on the simulation data that the authors used, the authors' method is best at identifying sets of genes that share a common pattern of differential expression (DE henceforth) across multiple experimental conditions. Please correct me if I'm wrong in my understanding. If I am correct on that, however, clusters of this type are constitutively different from the those that are identified by other clustering methods (such as k-means). It would be less confusing if the authors clarify*

this/these difference(s) from previous convex approaches.

Response: Yes, your understanding is correct. Please note that the K-means and SOM methods were applied to transformed data instead of the count data. There are two transformation steps, the log transformation and centralization, applied to the counts: the log transformation ensures that the gene expression is under logarithm scale, and the centralization removes the mean from the absolute magnitudes of gene expressions. Combining the two steps together, we are using K-means (as well as SOM) to cluster the log fold change (log-FC) of the expressions relative to the average expression of the gene. The log-FC, denoted by $\boldsymbol{\beta}_g$ in our context, characterizes the clusters in our MB clustering algorithm too. So genes identified as in the same cluster, either by K-means or by the MB method, share a similar pattern of differential expression across multiple treatments. We now clearly state this in the real data analysis: "..., upon 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."

The convexity in the shapes of the clusters detected by K-means is due to the symmetry of Gaussian distribution around the center and using the Euclidian distance. Due to the nature of RNA-seq data, the counts are discrete and skewed. We do not have a specific geometric pattern that we look for when separating the clusters.

5. *In addition, it would be helpful if they clearly stated what they thought were the advantages that their approach provided in contrast to previous methods.*

Response: Thank you for this suggestion. We now clearly listed our advantages over previous methods in the discussion section of the revised manuscript. We copy them here for your convenience:

" 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. Due 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 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 model-based hybrid-hierarchical clustering algorithm 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."

6. *In terms of the section describing the validation that was performed using real data, I think it would make for a stronger paper if the authors dug a little deeper into some of the clusters that their method identified. Specifically, I think it would be interesting if they could identify a few clusters that are unique to their method, i.e. they have low similarity to any of those that were identified with K-means (similarity could be measured via hyper-geometric distribution or the Jaccard index). Of those clusters that were missed by other clustering methods, it would be interesting for the authors to explore these some. For example, do they have a correlated expression profile? Do they share a common promoter motif? What GO function(s) do they share? Do they have any known co-associations (i.e. part of a metabolic pathway)? Any evidence of mutual interactions? The PlantCyc DB will be of help here, as may the STRING DB. I'm excited to see what you can turn up!*

   Response: Thanks for the specific hints to the approach of comparing two method via real data. We have done more extensive real data analysis, and added the new findings into section 5 on page 8. We copy the new paragraph here for your convenience:

   "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 3 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 more than 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."

We also included the Figure of clustering results from K-means and Model-based method using EM algorithm to help visualize the difference in the clusters identified by the two methods. This is now Figure 4 in the revised manuscript.

Response to Reviewer 3

Response: Thank you for careful review of our manuscript. We appreciate your insightful comments which help us substantially improve our manuscript.

Comments:

1. *It is a pity that more is not made of the application to real biological data, which would have made the paper significantly more interesting. Indeed, Supplementary Figure 5 , or a subset thereof, seems to make a better case for the improvement over k-means and SOM than Figure 3 does.*

   Response: Thank you for this comment. We have done more extensive real data analysis, and added the new findings into section 5 on page 8. We agree with you that the previous Supplementary Figure 5 makes a better case for the improvement over K-means. We followed your suggestion to include the Figure of clustering results from K-means and Model-based method using EM algorithm. This is now Figure 4 in the revised manuscript.

2. *It seems reasonable to presume that the improvements provided by MBCluster.Seq occur mostly for genes with low counts. Is this seen in practice? For example, in the simulation data, can it be shown that accuracy of assignment to clusters is dependent on expression level?*

   Response:

   It is true that the accuracy of assigning genes to clusters depends on expression level. As shown in Figure 1(b), 2(b), 3(b) and 4(b) in the Supplementary materials, the specificity, sensitivity and NMI scores for each clustering method all increases when the overall mean expression level α_g increases.

   However, when comparing across different clustering methods, the improvement of the model-based approach over K-means does *not* occur mostly for low reads as presumed. In our simulation study, we set all tuning parameters to be 1, and use the 7 true centers to initialized the clustering algorithms, then compare the result from MB to that from K-means. The genes are cut into six groups by their average read counts. In each group, the number (labeled as True# in the following table) of genes that are classified into their original groups and the the ratio (labeled as Ratio in the following table) of genes that are classified into their original groups are calculated for the two methods. This procedure is repeated 10 times and the results are averaged. Though the advantage of MB clustering is obvious, the following table show that the improvement is smaller (2.50% increase) when the average count is lower than 5, and the most improvement is for the genes with average counts between 5 and 10

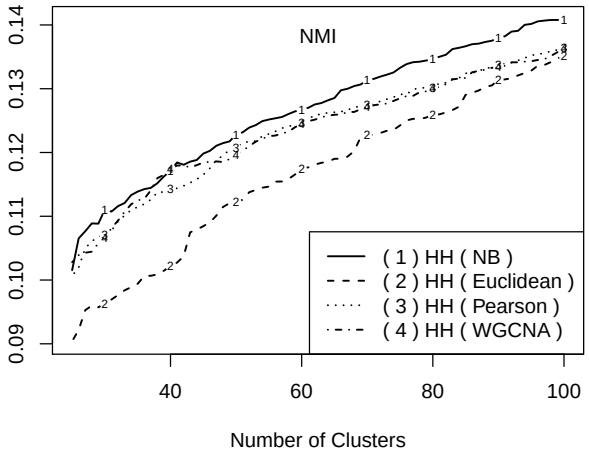
(4.44%). Hence with moderately small counts, the model-base method improves the most.

Group		K-Means		MB Cluster		Percent
Average Read	Total #	True #	Ratio	True #	Ratio	Improvement
≤ 5	47	28	0.601	29	0.616	2.50%
(5 , 10]	177	124	0.697	129	0.728	4.44%
(10 , 50]	2477	1985	0.801	2059	0.831	3.74%
(50 , 100]	2263	1902	0.840	1977	0.873	3.93%
(100 , 500]	4205	3618	0.860	3759	0.894	3.95%
> 500	831	729	0.877	756	0.910	3.76%

3. *I feel it would be useful to see a comparison between HH-NB and Pearson and euclidean distance based clustering, with the initial set of 200 clusters determined separately for each algorithm. This would more accurately reflect the practical application of the latter algorithms. Similarly, it would be useful to see the clustering(s) produced by WGCNA (Horvath et al.) included in the comparison, as it is a very similar clustering approach, which is commonly used, and which claims to improve upon Pearson correlation based hierarchical clustering.*

Response:

Thanks for the suggestion. We now performed K-means clustering with Euclidean, Pearson Correlation, Maximum distance, and compared these results to those using Model-Based hybrid-hierarchical clustering based on NB and Poisson models. The initial clusters are separately determined for each method. We also included the WGCNA method to build the hierarchical cluster using adjacency (similarity) function $a_{gg'} = |0.5 + \text{Corr}\left(\hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\beta}}_{g'}\right)|^{\gamma}$, where the power $\gamma = 10$ was selected by the *Scale-free Topology Criterion* as suggested. The advantage of HH-NB over other methods is still obvious as before. We have replaced Fig 3(b) in the revised manuscript with the following figure.



4. *Robinson and Oshlack is cited 3 times. The citation in sec 2.1 is correct (referring to scaling normalization) but citations in sec 1 and 2 should probably be to: Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.*

   Response: Thank you for this suggestion. It might be a better choice to cite the edgeR paper rather than the Robinson and Oshlack paper in sections 1 and 2. We have changed the citations accordingly.

5. *On page 7, it would be useful to clarify slightly that only K-means and SOM are applied to the log-RPKM transformed data, and that MBCluster.Seq was applied to count data. It should be obvious to readers and users that the distributional assumptions of the model would be violated if used on log-RPKM data. In a similar vein, what was the Pearson/euclidean distance clustering applied to - counts or log-RPKM values?*

   Response: Thanking you for pointing this out. We explicitly mention that the MBC method is applied to the count data now in section 5. We copy the sentence here for your convenience: "We also present results from the model-based clustering algorithms for the untransformed count data based on NB model.". Except when using the model-based clustering approach, all the other methods were applied to the log-transformed RPKM values.

6. *Finally, it would aid users of the R package if the authors were to consider expanding its documentation.*

   Response: Thank you for this suggestion. We are working on expanding the documentation of this package now.

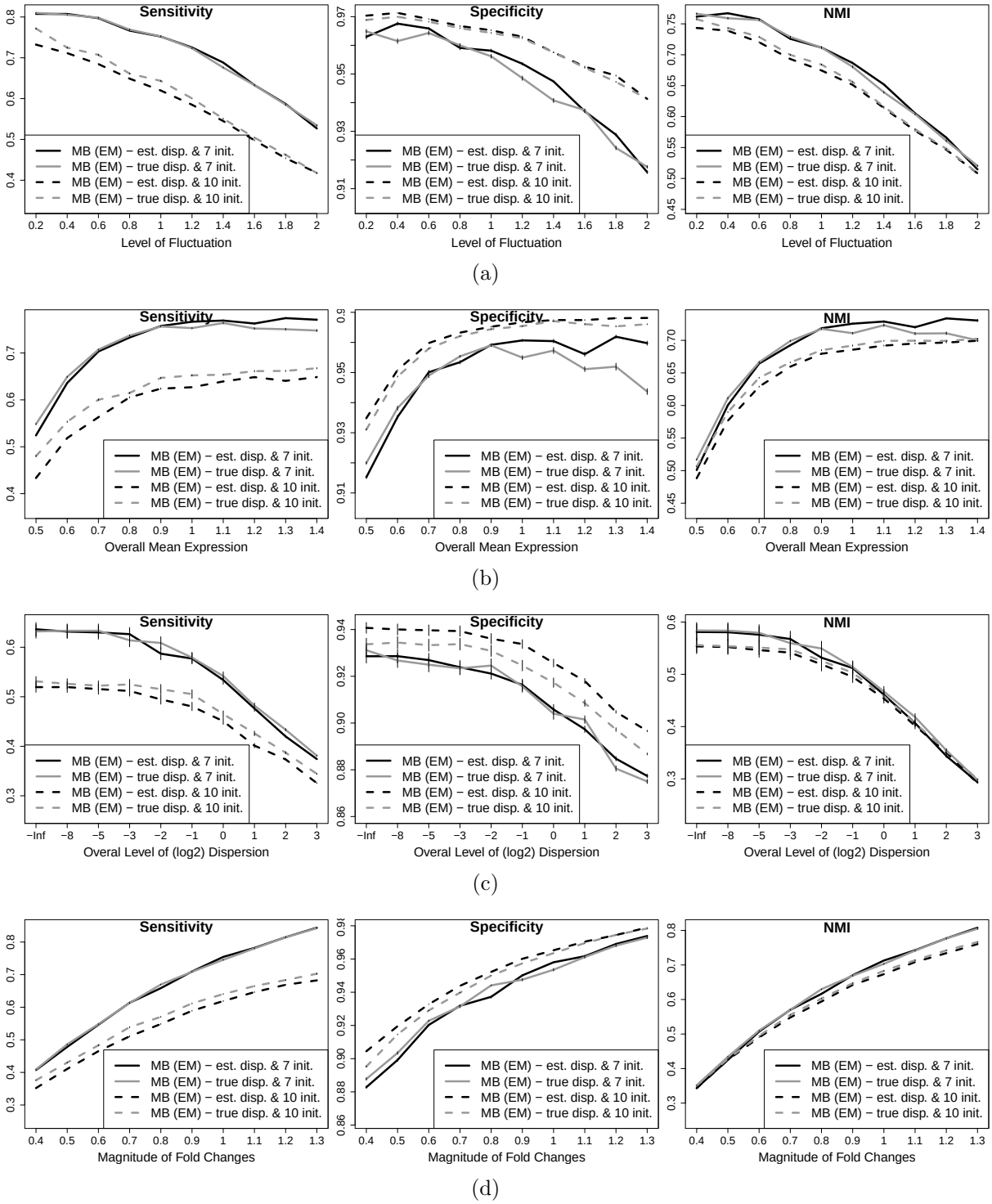
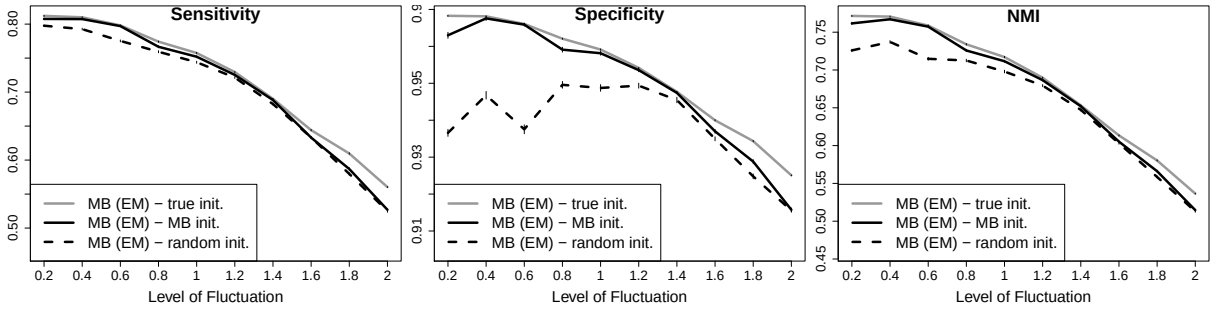
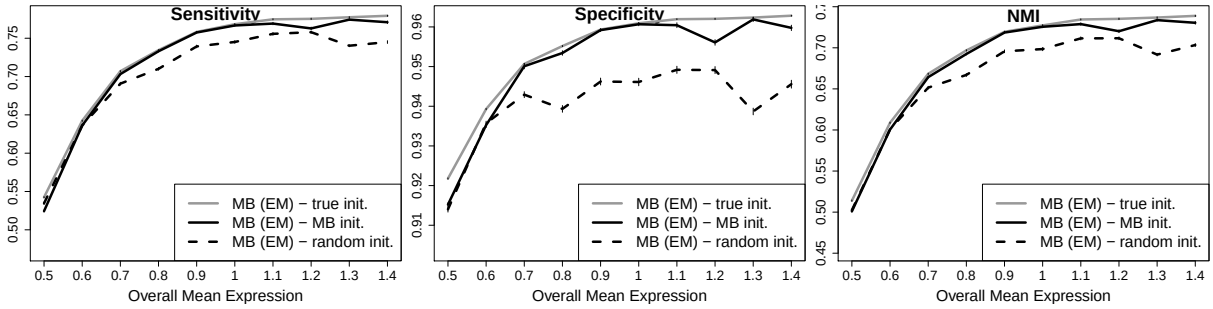


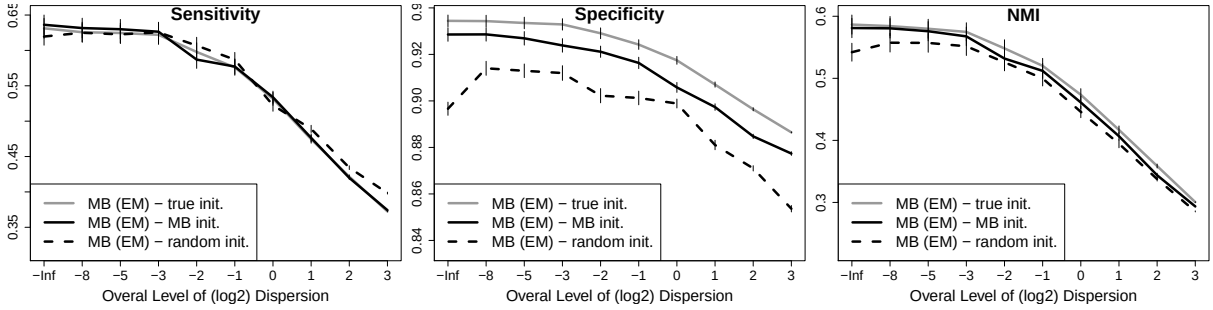
Figure 1: *Simulation Results: Estimation of dispersion parameters.* Clustering results from the MB-EM algorithms using true and estimated dispersion parameters were compared. For each parameter setting, result from 100 data sets were averaged and plotted on the line. The length of vertical bars represents standard error. Solid lines are for results of 7 clusters and dashed lines are for results of 10 clusters.



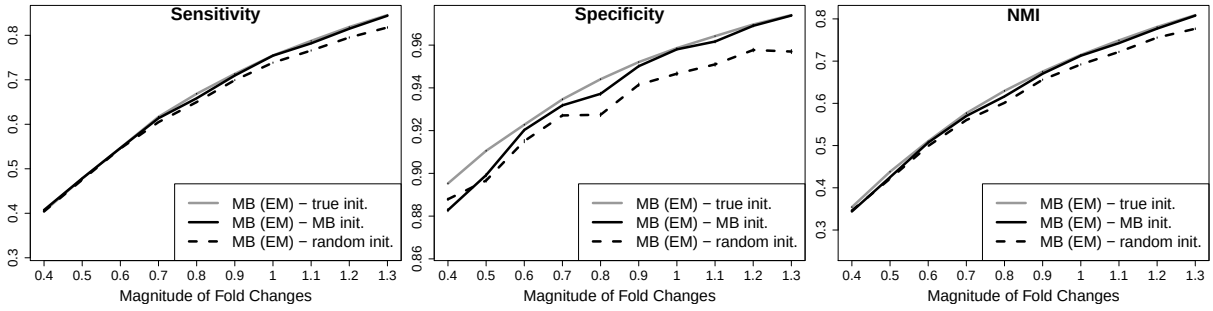
(a)



(b)

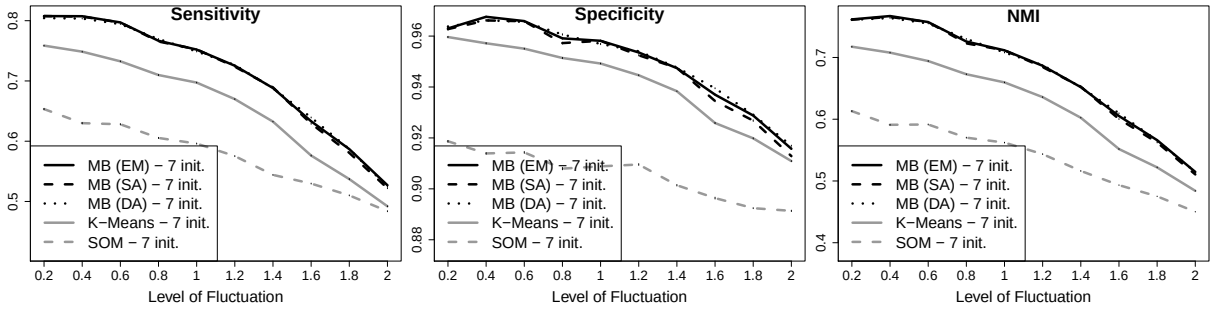


(c)

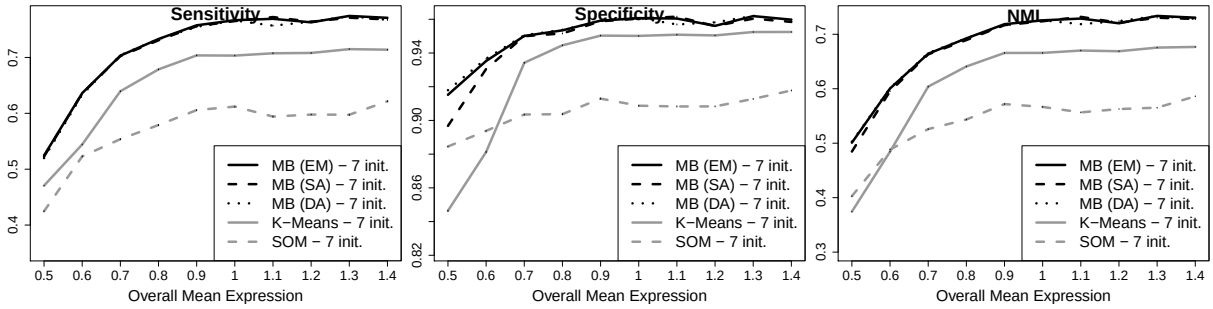


(d)

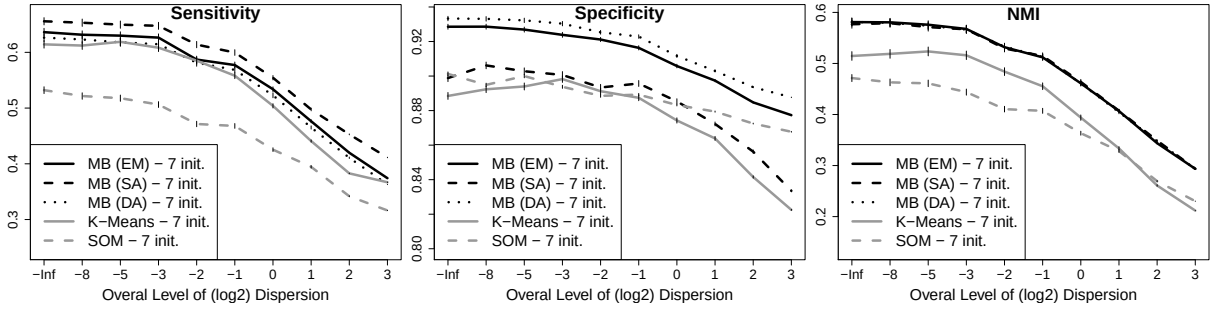
Figure 2: *Simulation Results: Evaluate initialization of cluster centers.* The two methods for initialization for MB-EM algorithms were compared: using initialization with model-based ALGORITHM 2 versus initialization with randomly picked objects(genes). For each parameter setting, results from 100 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.



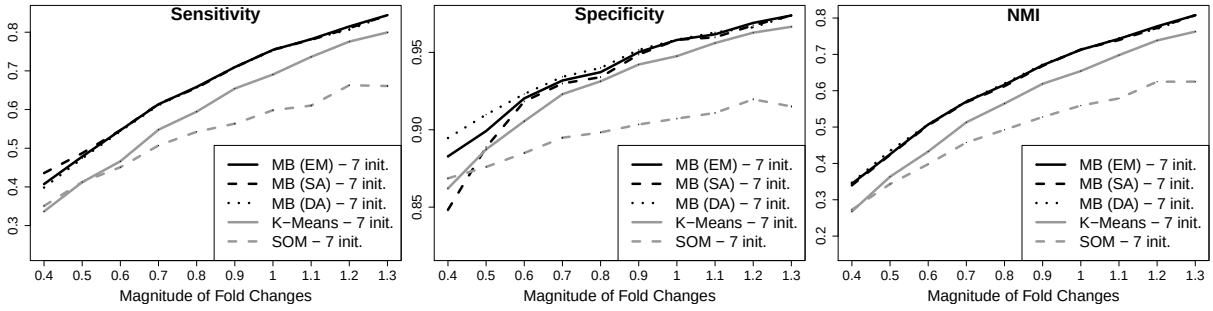
(a)



(b)



(c)



(d)

Figure 3: *Simulation Results: Results of seven clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms initialized by the same 7 cluster centers chosen by ALGORITHM 2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 100 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.

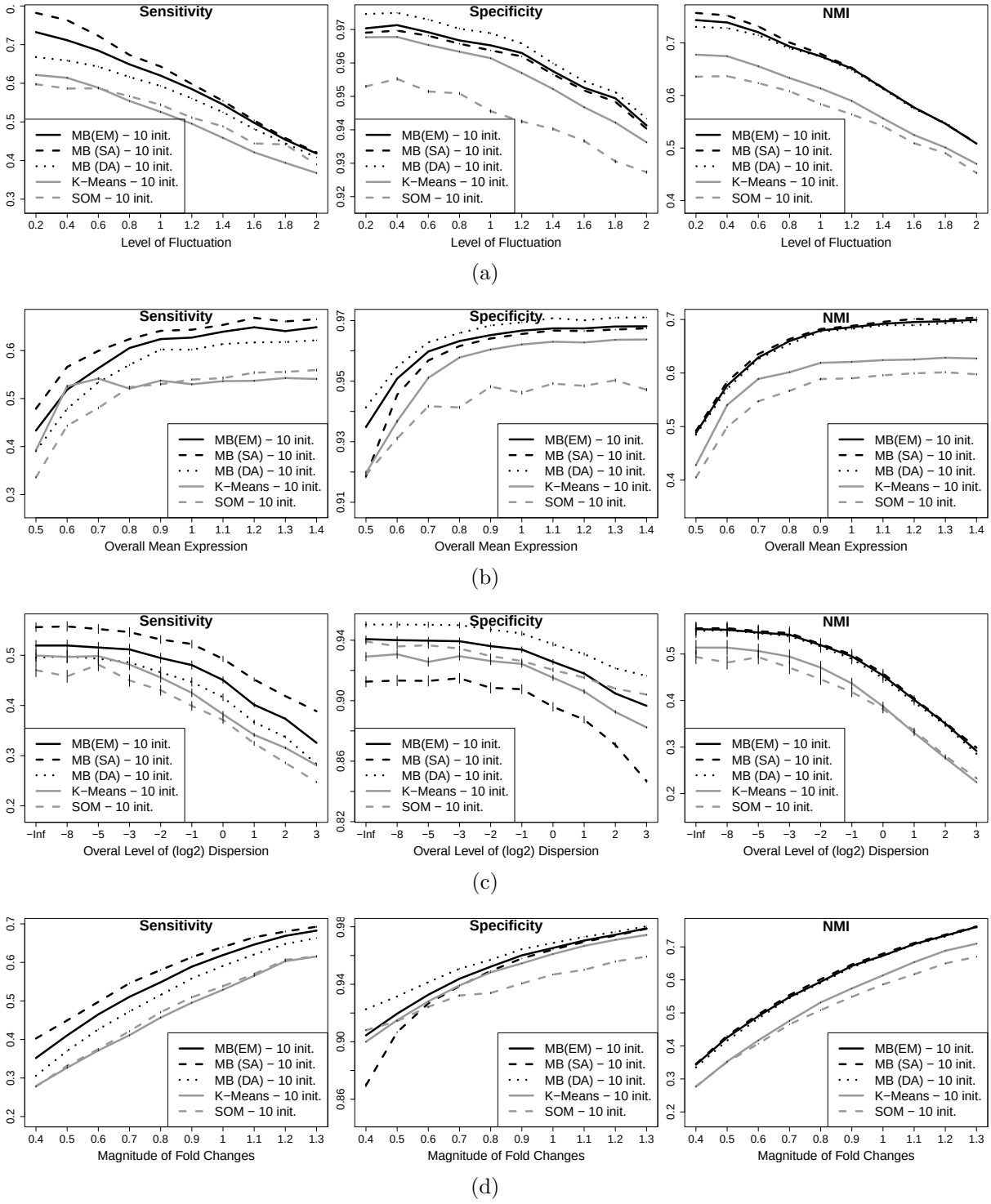


Figure 4: *Simulation Results: Results of ten clusters from different clustering methods.* The model-based methods include EM, DA and SA algorithms initialized by the same ten cluster centers chosen by ALGORITHM 2. The non-MB methods include the standard K-means and SOM. For each parameter setting, results from 100 data sets were averaged and plotted on the line. The length of each vertical bar represents standard error.