**Editor’s comments**

Thank you very much for submitting your Research Article entitled 'Clustering RNA-seq Expression Data using Grade of Membership Models' to PLOS Genetics. Your manuscript was fully evaluated at the editorial level and by independent peer reviewers. The reviewers appreciated the attention to an important problem, but raised some substantial concerns about the current manuscript. The reviewers would like you to clearly highlight the technical novelty of the work and/or novel biological insights that could be obtained from the method. Further, they have requested systematic comparisons to other related methods. Based on the reviews, we will not be able to accept this version of the manuscript, but we would be willing to review again a much-revised version. We cannot, of course, promise publication at that time.

Should you decide to revise the manuscript for further consideration here, your revisions should address the specific points made by each reviewer. We will also require a detailed list of your responses to the review comments and a description of the changes you have made in the manuscript.

If you decide to revise the manuscript for further consideration at PLOS Genetics, please aim to resubmit within the next 60 days, unless it will take extra time to address the concerns of the reviewers, in which case we would appreciate an expected resubmission date by email to [plosgenetics@plos.org](mailto:plosgenetics@plos.org" \t "_blank).

------------------------------------------------------------------------------------------------------------

Reviewer #1 comments

In this paper, the authors illustrate that grade of membership (GoM) models widely used in population genetics provide an attractive approach for clustering samples of RNA-seq data. The latent Dirichlet allocation model is applied to three RNA-seq gene expression data sets measured on either bulk samples or single cells and is compared with hierarchical clustering. The authors show that the GoM model can successfully highlight similarities among biologically-related tissues (or early embryonic development stages) and identify differentially expressed genes that recapitulate known biology.

It is a novel attempt to apply the latent Dirichlet allocation model and the Structure provides an effective visualization of the membership of each sample. The results reveal previously known biology underlying different cell types and developmental stages and batch effects. However, overall there needs to be more novelty either in the presented computational method (#1 below) or the biological results (#2 below). Lack of novelty diminished the reviewer’s enthusiasm.  
  
Major comments:

1. Lack of technical novelty. Sample clustering of RNA expression data is the area where

there are numerous existing methods. There needs to be either more technical advancement of the presented method (e.g., extending the latent Dirichlet allocation method). The authors could also perform or a more thorough comparison with alternative methods for clustering samples by soft assignment to justify the choice of the specific method used in the paper.

1. Lack of novelty in biological discovery. A demonstration of how the GoM model can reveal novel biology would improve the manuscript.

Minor comments:

1. Line 110: why K = 20? The authors should explain why K=20 was chosen.
2. Line 201-206: Jaitin et al, 2014. The authors refer to the figures in Jaitin et al. (2014): Figure 2A-B and Figure S7. The authors need to at least describe these figures in their figures so the readers do not have to find these figures, one being in the supplementary materials of Jaitin et al.

------------------------------------------------------------------------------------------------------------

Reviewer #2 comments

This manuscript discusses the application of Grade of Membership models (GoM) to RNAseq gene expression data, including the potential conceptual advantages and the results after applying the model to three gene expression datasets. They demonstrate that the GoM approach provides meaningful assignment of samples into (overlapping) groups. Additionally, they have a method for identifying the genes most relevant to each group, and thus provide a nice approach for interpreting gene expression structure and biological meaning. It is a nice discussion and clearly written but in its current form, lacking in novelty and deep analysis of the results/comparison to other approaches.

Major comments:

1. GoM models are related to a wide class of factorization methods, which also allow samples to represented as a combination of multiple overlapping clusters/factors, as the authors acknowledge. However they do not provide sufficient evidence/discussion of the relationship of GoM compared to these methods. The authors mention PCA couldn’t provide the same type of visualization/interpretation is this due to the fact that it is not sparse or not being able to interpret parameters as proportions? In particular, Sparse Factor Analysis, which the authors themselves have worked with, seems like a natural alternative that could capture many of the same effects (of course a transformation would have to be applied rather than working with raw counts) and interpretability. A direct comparison should be made, applying 1 or 2 matrix factorization methods to the same dataset(s) evaluated here, in terms of the enrichments found etc, and to demonstrate the advantages and disadvantages of the various approaches even simply regarding interpretation/visualization.
2. The GTEx results demonstrate that they capture clusters relevant to tissue biology, but are not particularly unexpected or detailed are they somehow better than standard clustering/PCA/SFA? The discussion of these results is quite long without providing clearly novel biological results or insight, or clear discussion of methodological advantage, though the visualization is nice.
3. Significance of the cluster enrichments and important genes is not clearly displayed/discussed, so it is hard to assess how meaningful it is.
4. The authors demonstrate application of the method to various data and capture different categories of effects notably, for one dataset they primarily identify batch effects but this raises an important point regarding the interpretation of the results overall that should be made more clear in the text that any cluster in any dataset could be technical or biological, and the method provides no guidance for distinguishing the two. That’s fine, and is true of the entire class of methods, but should be clearly stated. For instance, what if GTEx tissues had been confounded by batch? Even without confounding some of the clusters with the effect of interest, any clustering result are likely to include technical effects.
5. Some brief insight should be provided in the main text to explain their method for identifying the genes that characterize each cluster. Could this be applied to other methods than GoM?
6. Overall, the novelty in the manuscript has not been made fully clear it is an existing method is applied to 3 RNAseq datasets and the results are not biologically novel or discussed in much depth. The method is interesting, and is discussed conceptually but not in sufficient detail/comparison to other methods for readers to use this manuscript it as a guide when choosing methods for analysis

Minor comments:

1. In “Methods Overview” c\_{n+}” is not clearly defined. Assume it’s the number of reads in sample n?
2. The model is run multiple times and they select the result with the highest LL is this training LL? Since this is mostly an explanatory paper, the authors should clearly state the guarantees this approach provides in terms of optimization (ie no guarantee of a global optimum, the model is nonconvex, etc).
3. Statement that “modelbased approaches tend to be more efficient” should be supported with citation.
4. For Deng et al results, have they investigated role of batch or other technical variables, particularly when discussing results observed for particular samples/embryos?
5. Occasional inconsistency in capitalization of GTEX/GTEx

------------------------------------------------------------------------------------------------------------

Reviewer #3 comments

The authors proposing using admixture models to “soft” cluster samples from RNA-seq Expression experiments. They show that, for tissues samples, these methods reveal finer structure than just tissue-type. Additionally, in most cases, cluster-membership-profiles are more similar for tissues of the same type than for those differing. In addition, the authors applied this method to single-cell RNA-seq data; and show interesting heterogeneity for cells of the same cell-type. The paper is very clearly and concisely written, and the concepts are cleanly explained. The examples are also quite illustrative. This work has the potential to be broadly impactful, however I believe a deeper discussion of the downstream use and implications of results from the proposed method are needed.

Thoughts:

1. I would appreciate more discussion of the downstream use of this method. Do we gain something beyond searching for genes with heterogeneous expression within a cell-type? I think we do, because we get sub-clusters, and “soft” memberships, but I would like to see a discussion of what we might do with these. To me, this is a major missing component.
2. Is the additional clustering we see for cells within a cell type just chance heterogeneity (because we are using too large a K) — is there a way to show that this sub-clustering is “real”?
3. This method is very related to principle-component-analysis(PCA)/factor-analysis. To me, it appears to be identical, except there is no mean subtraction, the authors are using count data as the outcome, rather than continuous data — counts here are likely so large that they are basically continuous, though using a multinomial/poisson model accounts for mean-variance related heterogeneity. This is not a criticism of the paper — the authors are very clear that their goal is to put in conversation clustering methods and genomic problems — it is a connection I believe is worth noting in the paper though.
4. Hierarchical clustering with euclidean distance doesn’t take into account heterogeneity of variance of counts. This may be the driving factor in its failure here, which the authors note in the manuscript (though it also may not be). I would be interested to see how hierarchical clustering would perform if you use mahalanobis distance with a diagonal covariance matrix to account for the mean-variance relationship of a poisson; perhaps using a variance estimate of (average-number-of-reads-for-a-gene) + 1.