Reviewer 1 :

Major comments

...Overall there needs to be more novelty either in the presented computational method (#1 below) or the biological results (#2 below).

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.

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

We emphasise that the novelty here is in applying existing ideas ("admixture models" or "latent Dirichlet allocation") to a new domain - clustering and visualization of bulk and single-cell RNA-seq samples. As far as we are aware, no-one has previously produced visualizations anything like our Figures 1, 3 and 4 for any kind of expression data (either microarray or RNA-seq, bulk or single cell), and we believe they provide an innovative and informative summary of the structure in the expression data not provided by widely-used existing methods (see below). As some independent indirect support for this claim we note that our software for producing these visualizations has been downloaded 803 times (from 527 distinct IPs) since it was released in March 2016 (http://bioconductor.org/packages/stats/bioc/CountClust).

We were puzzled by the referee's suggestion to make "technical advancement", which seemed to be asking for "change for change's sake" rather than motivated by a specific deficiency of the method. However, we agree that a more thorough comparison with alternative methods would strengthen the paper, and so we have added these comparisons to the paper. Specifically we now include explicit comparisons with several widely-used methods for visualizing and clustering RNA-seq samples -- PCA, Multi-dimensional scaling, t-SNE, and hierarchical clustering; Figure 2, Supplementary Figures S2, S4 and S7 -- as well as sparse factor analysis (see below). We believe that the results clearly demonstrate the attraction of the GoM model results compared with these existing approaches: compare, for example, Figure 1 and Figure 2 (whole GTEx data), Figure1 and Supplementary Figure S4 (GTEx Brain data), Figure 4 and Supplementary Figure S7 (Deng et al single cell data). See Results for extensive discussion of the differences among methods.

In response to the second comment we have extended our analysis of the mouse preimplantation development single-cell RNA-seq data from Deng et al to include comparisons with previously-published similar data collected by qPCR. Our results clearly highlight an important issue, which is that the data from the qPCR and single-cell RNA-seq are not entirely concordant.

In addition, our new comparisons with existing methods demonstrate that the GoM model reveals much higher-resolution structure in expression profiles among brain tissues than existing methods, which essentially simply divide the expression patterns in brain tissues into two groups (cerebellar vs non-cerebellar).

While we would stop short of claiming these results reveal “novel biology” we do believe that our results on both these data and others demonstrate the potential for the GoM model to be useful in the analysis of data of this kind, and to be one tool in the toolbox that will ultimately lead to novel biological insights.

Minor comments:

1. **Line 110: why K = 20? The authors should explain why K=20 was chosen.**

As explained in the manuscript, we actually show results for K=5,10,15 and 20 (Supplementary Figure S1), and we focussed on K=20 because it shows the most high resolution structure. We also discussed choice of K in the Discussion - emphasising that there is no “correct” choice of K, and results with different K can complement one another. The choice of 20 as an upper bound in our analysis is for the simple practical reason that for larger K the Structure plot can be difficult to read due to the difficulty of selecting a suitable color palette; we have added a comment on this to the Discussion.

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.**

We removed the reference to their Figure S7, which was unnecessary, and perhaps confusing. We were not sure in what way the referee wanted us to expand on our current verbal description of Figures 2A-B (“the cluster structure evident in … their Fig 2A and 2B”).

Reviewer 2:

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.

In response to this comment, and also the comment of reviewer 1 calling for more comparisons with existing methods, we have added extensive comparisons with existing widely-used methods for analysis of expression data: PCA, multi-dimensional scaling, t-SNE and hierarchical clustering.

We have also added results for one type of Sparse Factor Analysis. However, we note that SFA is not (yet) a widely-used method in this context, and in fact applying it to RNA-seq data in this way raises several methodological issues: the reviewer mentions the transformation issue (which is non-trivial) but also there are *many* different ways to induce sparsity in factor analysis, and the question of whether to induce sparsity on the factors or the loadings. In addition, visualizing the results in an elegant way is also challenging, and we illustrate this in the new manuscript (Supplementary Figures S10, S11 and S12). We actually agree with the reviewer that SFA has considerable promise as a method for revealing structure in these kinds of data, but we believe a proper assessment of how it could and should be best used in this regard would be a full research project in itself.

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.

We now provided a more direct demonstration and discussion of the benefits of the GoM model results compared to standard (hierarchical) clustering/PCA and t-SNE for the GTEX data. See particularly Figure 1 vs Figure 2 (for the whole data) and Figure 1 vs Supplementary Figure S4 for the brain data.

3. Significance of the cluster enrichments and important genes is not clearly displayed/discussed, so it is hard to assess how meaningful it is.

We have added the significance values (P-values) of the GO terms that are associated with the important genes to the results of our cluster enrichment analysis (Gene Ontology; see Tables 1-3 and Supplemental Table 4).

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.

We agree, and added the following paragraph about the GTEx results in response to this comment:

Although it is not surprising that samples cluster by tissue, other results could have occurred. For example, samples could have clustered according to technical variables, such as sequencing batch or sample collection center. While our results do not exclude the possibility that technical variables could have influenced these data, the t -SNE and GoM results clearly demonstrate that tissue of origin is the primary source of heterogeneity, and provide a useful initial assurance of data quality.

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?

We added an additional clarification on this in the main text (at the end of Methods overview), to supplement our full description in the Methods section. Although we made a specific choice of measure here based on the use of KL divergence, we would expect other choices also to work well, so we don't overemphasise this issue.

Regarding whether this could be applied to other methods -- yes, if those methods provide an explicit estimate of expression level in each cluster, but not otherwise. (One limitation of the PCA, multidimensional scaling, and t-SNE results is that they do not provide an explicit clustering of samples, only a visual representation from which clusters can sometimes be picked out by eye. The lack of explicit clusters makes it impossible to provide cluster annotations, a point we now make in our discussion of the GTEx results.)

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

We believe that with the addition of the more in-depth comparisons with existing methods, our revised manuscript now provides a much more effective guide to readers who wish to choose among methods for analysis. (See also the response to reviewer 1 regarding biological novelty).

Reviewer 3:

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.

In our view, there are so many different ways to follow up on results of an exploratory analysis tool like this that we are cautious about highlighting any particular one. However, we have expanded on our discussion of the results on the mouse preimplantation data from Deng et al, and made a fuller comparison with previous results on similar data measured by qPCR (check Supplementary Figures S8 and S9). More generally we believe that our examples demonstrate a wide breadth of possible applications, and we hope that the reviewer finds that the comparisons with existing methods that we have added to our revised manuscript provide a better indication of the benefits of the method compared with the original submission.

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"?

The general question of how to determine the "significance" of observed structure is tricky, and is one faced by all the methods we consider here. (Eg if a cluster shows up in a PCA or tSNE plot, is it "real"?) We do provide some general discussion on choice of K, that we hope readers will find helpful.

However, for the specific case of the Deng data considered here, the results of the GoM analysis (and indeed other methods, like PCA) actually show *less* substructure than might be expected based on known biology and previous results on similar data generated by qPCR. We have added extensive discussion of this issue and explicit comparison with previous data.

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.

We added a paragraph about the relationship between GoM models, factor analysis and PCA to the discussion - specifically, all of these can be viewed as matrix factorization methods, but with different constraints on the matrices. However, as we hope our new comparisons with PCA (and SFA) demonstrate, these methods provide very different results and visualizations of RNA-seq data.

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.

We added Supplementary Figure S3 to show the effects of standardization on hierarchical clustering. For all the standardizations we tried (including the one suggested by the referee) the GoM model provided more accurate results.